Queueing Syst (2017) 86:1–33
DOI 10.1007/s11134-017-9516-3

Fluid and diffusion approximations of probabilistic matching systems

Burak Büke[1] · **Hanyi Chen**[1]

Received: 12 January 2015 / Revised: 21 January 2017 / Published online: 23 February 2017

Abstract This paper focuses on probabilistic matching systems where two classes of users arrive at the system to match with users from the other class. The users are selective and the matchings occur probabilistically. Recently, Markov chain models were proposed to analyze these systems; however, an exact analysis of these models to completely characterize the performance is not possible due to the probabilistic matching structure. In this work, we propose approximation methods based on fluid and diffusion limits using different scalings. We analyze the basic properties of these approximations and show that some performance measures are insensitive to the matching probability, agreeing with the existing results. We also perform numerical experiments with our approximations to gain insight into probabilistic matching systems.

Keywords Matching systems · Fluid approximations · Diffusion approximations

Mathematics Subject Classification 60B10 · 60K25 · 90B22

1 Introduction

The Internet has provided society a new medium to carry out business and personal transactions. In this work, our goal is to provide tractable methods to analyze probabilistic matching systems introduced in Büke and Chen [3] to study the web portals that

✉ Hanyi Chen
H.Chen-29@sms.ed.ac.uk

Burak Büke
B.Buke@ed.ac.uk

[1] School of Mathematics, The University of Edinburgh, King's Buildings, James C. Maxwell Building, Peter Guthrie Tait Road, Edinburgh EH9 3FD, UK

serve as a meeting point for suppliers and customers of a specific product or service. The examples of such systems include employment and rental portals, matrimonial and dating Web sites and general purpose classified advertisement Web sites.

The users of a probabilistic matching system can be classified into two groups as customers (for example, employers) and suppliers (for example, employees). Customers arrive at the system according to a stochastic process. When a customer arrives at the system, she searches the list of suppliers to see whether there is anybody selling the product (or the service) she demands. If she finds suppliers with suitable products, she buys a product, choosing one uniformly at random, and both the customer and the supplier leave the system together. If there are no suitable products available, then she posts an advertisement on the system indicating her demand and waits until a supplier with a suitable product arrives at the system. The suppliers also exhibit a similar behavior.

The double-ended queue introduced in Kashyap [12] is a precursor for the matching system and considers the queueing process of taxis and customers at a taxi stop. Taxis and customers arrive at the stop according to independent Poisson processes and if a taxi (customer) arrives when there are no customers (taxis) waiting at the stop, she waits until a customer (taxi) arrives. Recently, there has been a growing interest to study matching systems which can be perceived as generalizations of double-ended queues. For these systems, each class of users has several subclasses, which we refer to as types, and these types determine whether users from different classes can match or not. Drawing an analogy between these matching systems and the taxi problem of Kashyap [12], in these systems there are different types of taxis, each of which serves a set of neighborhoods, and a taxi accepts a customer, in other words matches with the customer, if and only if she is going to a neighborhood served by the type that the taxi belongs to. For these models, once the types of users are known, the matchings occur deterministically and the main goal is to devise policies to decide on which users should be matched with each other. Caldentey et al. [5] introduce a matching system with two classes of users, namely customers and servers, where each class has several types. The types of servers with which a customer of a given type can match are determined using a bipartite graph. The model can be thought of as a discrete-time process where exactly one customer and one server arrive at each time period. The types of the arriving customer and server are independent and follow a given probability measure. If there are users who can match after arrivals occur, they are matched on a first-come-first-served basis. The authors conjecture necessary and sufficient conditions on the probability measure for the stability of these systems, and they prove that if a given system is stable the matching rates of different types converge to a limit. They also study stability of some simple systems. Adan and Weiss [1] prove that the conditions conjectured in [5] are necessary and sufficient and they prove that the stationary probabilities have a product form. Bušić et al. [4] generalize this model by dropping the independence of arriving types. They consider matching policies other than first-come-first-served. Using the fact that the conditions conjectured in [5] are necessary for stability, they show that matching the longest queue has a maximal stability region, i.e., the system is stable for any probability measure that satisfies the necessary conditions. They also show that matching the shortest and some priority policies does not have a maximal stability region and prove some sufficient conditions

for the stability of these matching systems. Mairesse and Moyal [15] generalize the bipartite matching model and develop necessary conditions for matching networks with general topology.

In a recent work, Gurvich and Ward [10] study a system where the matchings do not need to be pairwise, but more than two users can match based on their classes. The arrivals to each class occur according to a continuous-time stochastic process, and the system controller decides who will match with whom. Gurvich and Ward [10] consider the objective of minimizing the finite horizon inventory holding cost and suggest a periodic review policy which relies on solving a linear program. They show that as arrival rates increase to infinity, the suggested periodic review policy is asymptotically optimal.

The key feature which differentiates probabilistic matching systems from the conventional matching systems in the literature is the probabilistic nature of the matching process. When a customer arrives at the system, she checks the products of all the suppliers and may find each product suitable with a given probability independent of the others. Hence, with positive probability she may not find any suitable product, even if there are several suppliers offering a product in the system. To make this argument more concrete, consider an employment portal as an example. An employer arriving at the employment portal first scans through the resumés of all the employees in the system, and she may hire each potential employee with a given probability. There is a positive probability that she may not find any of the existing candidates suitable, in which case she posts a job advert and waits in the system until a suitable candidate arrives. Hence, unlike the double-ended queues, users from different classes can coexist in the system when the matchings are probabilistic, which makes it essential to model the queueing system as a two-dimensional stochastic process. Büke and Chen [3] study the effects of the matching probability on the performance of these systems using an exact analysis and show that if uncontrolled these systems are unstable. Büke and Chen [3] also suggest stabilizing admission control policies to decide when to accept an arriving user into the system based on the system size and analyze some performance measures (for example, throughput and average queue length) under the suggested policies.

As indicated by the employment portal example, probabilistic matching systems are especially useful when the operator does not have any control over the matching behavior. For example, in an employment portal, employers have the full list of employees and can decide to hire one of the candidates irrespective of their queueing behavior, for example, the operator cannot force an employer to hire employees based on the their arrival time. Dating and matrimonial Web sites, rental or auto trade portals are examples of other systems where the matching cannot be controlled by the operator. The probabilistic matching behavior complicates the analysis of these systems and renders a complete exact analysis intractable. Hence, in this work we propose approximation methods based on fluid and diffusion limits under two different scalings. Under our first scaling, we only scale time and space and keep the matching probability constant to obtain the limiting processes. We show that under this scaling both fluid and diffusion limits do not depend on the matching probability, which implies that the users from at most one class accumulate in the system and the probability of a user finding a match upon arrival approaches either zero or one.

In many applications, it is crucial to study the effect of the matching probability on the system performance. To provide tools which address the matching probability explicitly, we propose a second scaling that also handles the abandonment of impatient users and scales the matching probability and the abandonment rate along with the time and space. The resulting fluid and diffusion limits under this scaling involve differential equations which are not tractable analytically in the general case, although we can derive an analytical formula for the fluid limit when there are no abandonments. Büke and Chen [3] show that some performance measures, such as the difference between the average queue lengths of different classes, are insensitive to the matching probability under certain control policies. Despite not imposing any control policy, similar to the results in [3] we show that the difference between queue lengths for different classes is also insensitive to the matching probability in the fluid limit.

In addition, we analyze the asymptotic behavior of the fluid limits. We first compare the fluid limits under both scalings, i.e., limits with and without scaling the matching probability, and show that when the abandonment rate is zero, the fluid limits in both scaling regimes agree with each other as time goes to infinity. Further, we show that for nonzero abandonment rates, the fluid limits converge to a unique fixed point, which is representative of the long-run average number of users in the system. We prove that as the abandonment rate increases, the fixed point component for the class with lower arrival rate first experiences an increase and then decrease, while for the class with higher arrival rate it decreases monotonically. Finally, we present numerical results to understand the approximation quality of probabilistic matching systems by the fluid limits. We also analyze the properties of fluid and diffusion limits in the second scaling regime using numerical methods.

There exists an extensive literature on fluid and diffusion approximations for Markovian systems with abandonments. Ward and Glynn [18] suggest diffusion approximations for the $M/M/1$ queue with exponential abandonments. They generalize these results to arrival, service and abandonment times with general distributions in [19]. Garnett et al. [9] consider the $M/M/N$ queue with exponential abandonments and suggest diffusion approximations under the Halfin–Whitt regime (see Halfin and Whitt [11]). Generalizing these results, Dai and He [7] and Mandelbaum and Momčilović [16] suggest diffusion approximations for many-server queues with general arrival, service and abandonment times. A recent work by Liu et al. [14] suggests diffusion approximations for the double-ended queue where arrivals are renewal processes and customers abandon the system if they cannot find a match after an exponential time. This paper is closest to our work in nature and the scaling they consider has similarities with both scalings presented here. In [14], the matching probability is not considered explicitly and is fixed to one, which is similar to the scaling we present in Sect. 3, whereas the abandonment rate is scaled to go to zero, similarly to our second scaling in Sect. 4. Even though we restrict ourselves to Poisson arrival processes, our work extends [14] by assuming probabilistic matching structure.

1.1 Notation

We assume that all stochastic processes used in this paper are defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The sample paths of the stochastic processes are assumed to take values in $\mathbb{D}[0, \infty)$, i.e., the space of right continuous functions with left limits. $X(\omega, t)$ denotes the value of stochastic process X at time t for sample point ω. We suppress the ω in the notation if we do not need to refer to the sample point explicitly. In this work, our goal is to develop approximations for probabilistic matching systems using convergence of probability measures. We use $\xrightarrow{\text{a.s.}}$, $\xrightarrow{\mathbb{P}}$ and $\Rightarrow$ to denote almost sure convergence, convergence in probability and convergence in measure, respectively. We also occasionally need to use the indicator function $\mathbb{I}_\rho$ which takes the value 1 if the proposition ρ is true and 0 otherwise. We use $\mathbb{N}$ and $\mathbb{R}_{\geq 0}$ to denote the sets of nonnegative integers and nonnegative real numbers, respectively.

2 The probabilistic matching model

In this work, we study probabilistic matching systems introduced in Büke and Chen [3], where two classes of users, indexed by $i = 1, 2$, arrive at the system to be matched with users of the other class. We assume that class-i users arrive according to a Poisson process with rate λ_i. Any given pair of class-1 and class-2 users can match with each other with probability q independent of other users. Let $X_i(t)$ be the number of class-i users in the system at time t. When a class-1 user arrives at time t, she checks the class-2 queue to see whether there exist any suitable users that she can match with. If she can find one or more suitable class-2 users to match with, she chooses one of them uniformly at random and they leave the system together. Otherwise, she joins the class-1 queue and waits in the system until she is picked by an arriving class-2 user. Due to the independence of matchings, a class-1 user finds a suitable class-2 user to match upon arrival with probability $1 - (1 - q)^{X_2(t)}$ and is not able to match with anyone with probability $(1 - q)^{X_2(t)}$. The same rules also apply when class-2 users arrive. For the analysis in Sect. 4, we also assume that the users are impatient and each user abandons the system without being matched after waiting an exponential time with rate $\gamma \geq 0$. For notational convenience, in the remainder of this paper we also assume that the system under consideration is initially empty, i.e., $(X_1(0), X_2(0)) = (0, 0)$ with probability one.

Under the assumption of Poisson arrivals, the number of users in a probabilistic matching system, $\{(X_1(t), X_2(t)), t \geq 0\}$, can be modeled as a continuous-time Markov chain (CTMC) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the generator matrix

$$Q_{(i,j)(l,k)} = \begin{cases} \lambda_1(1-q)^j, & \text{if } l = i+1 \text{ and } k = j, \\ \lambda_2(1-q)^i, & \text{if } l = i \text{ and } k = j+1, \\ \lambda_1\left(1-(1-q)^j\right) + \gamma j, & \text{if } l = i \text{ and } k = j-1 \geq 0, \\ \lambda_2\left(1-(1-q)^i\right) + \gamma i, & \text{if } l = i-1 \geq 0 \text{ and } k = j, \\ -(\lambda_1 + \lambda_2 + \gamma(i+j)), & \text{if } l = i \text{ and } k = j, \\ 0, & \text{otherwise.} \end{cases}$$

The above model reduces to the one introduced in [3] if users do not abandon the system ($\gamma = 0$).

It is sometimes useful in our analysis to express the queue length processes, $X_i(t)$, as the difference of counting processes. We define $A_i(t)$ and $R_i(t)$ to be the number of arrivals and the number of user abandonments from class-i up to time t, respectively. Similarly, defining $M(t)$ to be the number of matched pairs up to time t, we have the basic relation

$$X_i(t) = A_i(t) - M(t) - R_i(t) \quad \text{for all} \quad t \geq 0 \quad \text{and} \quad i = 1, 2.$$

The essential element distinguishing a probabilistic matching system from a conventional queuing system is the matching probability q. To see this, consider a probabilistic matching system with no abandonments ($\gamma = 0$). For systems with matching probability $q = 1$, class-1 and class-2 users cannot coexist in the system at any time. Hence, the probabilistic matching system can be modeled as a continuous-time random walk on the integers $\{X(t), t \geq 0\}$, where $X(t) = k$ if $(X_1(t), X_2(t)) = (0, k)$ and $X(t) = -k$ if $(X_1(t), X_2(t)) = (0, k)$. Also when $q = 1$, the number of matched pairs up to time t is equal to the minimum of class-1 and class-2 arrivals. Hence,

$$X_i(t) = A_i(t) - M(t) = A_i(t) - \min\{A_1(t), A_2(t)\}, \quad \text{for all } t \geq 0 \quad \text{and} \quad i = 1, 2.$$

However, when $0 < q < 1$, analyzing the matching process $M(t)$ is far more difficult. The one-dimensional distribution of the matching process, $\mathbb{P}(M(t) = k)$ for a given $t \geq 0$ and $k \in \mathbb{N}$ is provided in [3], and its complicated nature indicates the difficulty in fully characterizing the law of the matching process. Hence, in this paper we propose fluid and diffusion approximations for probabilistic matching systems.

3 Fluid and diffusion approximations with constant matching probability

In this section, we focus on fluid and diffusion approximations for probabilistic matching systems obtained by only scaling time (or equivalently the arrival rates) and space while keeping the matching probability constant. For scalings with a constant matching probability, we assume that the users do not abandon the system without being matched, i.e., $\gamma = 0$. Both fluid and diffusion limits under this scaling fail to represent the matching probability explicitly, indicating the need to scale the matching probability as studied in Sect. 4.

3.1 Fluid limits

We start by defining the scaled process $\{(\bar{X}_1^n(t), \bar{X}_2^n(t)), t \geq 0\}$ as

$$\bar{X}_i^n(t) = \frac{X_i(nt)}{n}, \quad i = 1, 2, \quad \forall t \geq 0.$$

We derive the limiting process of $\{\bar{X}_i^n(t), t \geq 0\}$ as $n \to \infty$. For any $\omega \in \Omega$, we say that $\bar{X}_i^n(\omega, \cdot)$ converges uniformly on compact sets (u.o.c.) to $\bar{X}_i(\omega, \cdot)$ if $\sup_{0\leq t\leq T} |\bar{X}_i^n(\omega, t) - \bar{X}_i(\omega, t)|$ converges to 0 for all $T > 0$ as $n \to \infty$. A direct application of the functional strong law of large numbers (see, for example, [2,6,20]) to Poisson arrival processes yields

$$\bar{A}_i^n(t) := \frac{A_i(nt)}{n} \xrightarrow{\text{a.s.}} \lambda_i t \text{ u.o.c.} \quad \text{as } n \to \infty, \quad i = 1, 2. \tag{1}$$

As users of a class accumulate in the system, the users of the other class are more likely to match upon their arrival. This implies that class-1 and class-2 users are unlikely to accumulate in the system at the same time. Lemma 1 formalizes this argument.

Lemma 1 *For any fixed* $k > 0$, $\min\{\frac{X_1(nt)}{n^k}, \frac{X_2(nt)}{n^k}\} \xrightarrow{\text{a.s.}} 0$ u.o.c. *as* $n \to \infty$.

Proof If $q = 1$, since class-1 and class-2 do not coexist in the system, for any $t \geq 0$, $\min\{X_1^n(t), X_2^n(t)\} = 0$, and hence the desired conclusion follows trivially. If $0 < q < 1$, to simplify the notation, define $I^{n,k}(t) := \min(\frac{X_1(nt)}{n^k}, \frac{X_2(nt)}{n^k})$. The Borel–Cantelli lemma implies that $I^{n,k} \xrightarrow{\text{a.s.}} 0$ u.o.c. if for any $T > 0$ and $\epsilon > 0$

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\sup_{0\leq t\leq T} I^{n,k}(t) \geq \epsilon \right) < \infty.$$

Choosing $a \in (0, k)$ and $N \geq 2$ such that $N^{-a} < \epsilon$, we have

$$\begin{aligned} \sum_{n=1}^{\infty} \mathbb{P}\left(\sup_{0\leq t\leq T} I^{n,k}(t) > \epsilon \right) \leq \sum_{n=1}^{N-1} \mathbb{P}\left(\sup_{0\leq t\leq T} I^{n,k}(t) > \epsilon \right) \\ + \sum_{n=N}^{\infty} \mathbb{P}\left(\sup_{0\leq t\leq T} I^{n,k}(t) \geq n^{-a} \right). \end{aligned} \tag{2}$$

We will now prove that the right-hand side of (2) converges. We take $\lambda = \lambda_1 + \lambda_2$, and for any $m \in \mathbb{N}$, we have

$$\begin{aligned} &\mathbb{P}\left(\sup_{0\leq t\leq \frac{m+1}{n^2}} I^{n,k}(t) \geq n^{-a} \,\middle|\, \sup_{0\leq t\leq \frac{m}{n^2}} I^{n,k}(t) < n^{-a} \right) \\ &= \mathbb{P}\left(\sup_{\frac{m}{n^2}\leq t\leq \frac{m+1}{n^2}} I^{n,k}(t) \geq n^{-a} \,\middle|\, \sup_{0\leq t\leq \frac{m}{n^2}} I^{n,k}(t) < n^{-a} \right) \\ &= \mathbb{P}\left(\sup_{\frac{m}{n^2}\leq t\leq \frac{m+1}{n^2}} \min(X_1(nt), X_2(nt)) \geq n^{k-a} \,\middle|\, \sup_{0\leq t\leq \frac{m}{n^2}} \min(X_1(nt), X_2(nt)) < n^{k-a} \right) \end{aligned}$$

$$\leq \sum_{j=0}^{\infty} \frac{e^{-\frac{\lambda}{n}}(\frac{\lambda}{n})^j}{j!} j(1-q)^{n^{k-a}}$$
$$= \frac{\lambda}{n}(1-q)^{n^{k-a}}. \tag{3}$$

We see that the inequality in (3) holds using the following argument. For both $X_1(nt)$ and $X_2(nt)$ to reach a level above n^{k-a} at some point during $[\frac{m}{n^2}, \frac{m+1}{n^2}]$, at least one of the arrivals occurring during $[\frac{m}{n^2}, \frac{m+1}{n^2}]$ should fail to match and stay in the system upon arrival when facing at least $\lfloor n^{k-a} \rfloor$ users from the other user queue (where $\lfloor x \rfloor$ is the smallest integer no smaller then x). If we observe j arrivals during this time interval, the probability of this event is bounded by $j(1-q)^{n^{k-a}}$. Then, for any fixed $T > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{0\leq t\leq T} I^{n,k}(t) \geq n^{-a}\right) \\
&= \mathbb{P}\left(\sup_{0\leq t\leq T} I^{n,k}(t) \geq n^{-a} \,\middle|\, \sup_{0\leq t\leq T-\frac{1}{n^2}} I^{n,k}(t) < n^{-a}\right) \mathbb{P}\left(\sup_{0\leq t\leq T-\frac{1}{n^2}} I^{n,k}(t) < n^{-a}\right) \\
&\quad + \mathbb{P}\left(\sup_{0\leq t\leq T} I^{n,k}(t) \geq n^{-a} \,\middle|\, \sup_{0\leq t\leq T-\frac{1}{n^2}} I^{n,k}(t) \geq n^{-a}\right) \mathbb{P}\left(\sup_{0\leq t\leq T-\frac{1}{n^2}} I^{n,k}(t) \geq n^{-a}\right) \\
&\leq \mathbb{P}\left(\sup_{0\leq t\leq T} I^{n,k}(t) \geq n^{-a} \,\middle|\, \sup_{0\leq t\leq T-\frac{1}{n^2}} I^{n,k}(t) < n^{-a}\right) \\
&\quad + \mathbb{P}\left(\sup_{0\leq t\leq T-\frac{1}{n^2}} I^{n,k}(t) \geq n^{-a}\right) \\
&\leq \sum_{m=0}^{Tn^2} \mathbb{P}\left(\sup_{0\leq t\leq \frac{m+1}{n^2}} I^{n,k}(t) \geq n^{-a} \,\middle|\, \sup_{0\leq t\leq \frac{m}{n^2}} I^{n,k}(t) < n^{-a}\right).
\end{aligned}$$

Hence, using (3), we conclude

$$\mathbb{P}\left(\sup_{0\leq t\leq T} I^{n,k}(t) \geq n^{-a}\right) \leq \sum_{m=0}^{Tn^2} \frac{\lambda}{n}(1-q)^{n^{k-a}} = T\lambda n(1-q)^{n^{k-a}}.$$

Using simple calculus, we can show that there exists an N_a such that for all $n > N_a$ we have $(1-q)^{n^{k-a}} < n^{-3}$, which implies

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\sup_{0\leq t\leq T} I^{n,k}(t) \geq n^{-a}\right) \leq T\lambda \sum_{n=0}^{\infty} n(1-q)^{n^{k-a}} < \infty.$$

Hence, the left-hand side of (2) converges and the result follows. □

Theorem 2 $\bar{X}_i^n \xrightarrow{\text{a.s.}} \bar{X}_i$ u.o.c. as $n \to \infty$, *where*

$$\bar{X}_i(t) = \lambda_i t - \min\{\lambda_1, \lambda_2\}t, \quad i = 1, 2. \tag{4}$$

Proof Equation (1) and Lemma 1 imply that there exists a $\Omega' \subset \Omega$ with $\mathbb{P}(\Omega') = 1$ where, for every $\omega \in \Omega'$,

$$\frac{A_i(\omega, nt)}{n} \to \lambda_i t \text{ u.o.c.} \quad \text{for } i = 1, 2, \tag{5}$$

$$\min\left\{\frac{X_1(\omega, nt)}{n}, \frac{X_2(\omega, nt)}{n}\right\} \to 0 \text{ u.o.c.} \tag{6}$$

Our first goal is to show

$$\bar{M}^n(\omega, t) := \frac{M(\omega, nt)}{n} \to \min\{\lambda_1 t, \lambda_2 t\} \text{ u.o.c.} \quad \text{as } n \to \infty \quad \text{for all } \omega \in \Omega'. \tag{7}$$

Without loss of generality, assume $\lambda_1 \geq \lambda_2$ and suppose that there exists an $\omega' \in \Omega'$ for which (7) does not hold, i.e., we can find $T > 0, \delta > 0$ and a sequence n_j such that

$$\lim_{j\to\infty} \sup_{0\leq t\leq T} \left| \frac{M\left(\omega', n_j t\right)}{n_j} - \lambda_2 t \right| > \delta.$$

Using (5) and the fact that $M(\omega', t) \leq \min\{A_1(\omega', t), A_2(\omega', t)\}$ for all $t \geq 0$, this implies that there exists a sequence t_j such that $0 \leq t_j \leq T$ and

$$\lim_{j\to\infty} \lambda_2 t_j - \frac{M(\omega', n_j t_j)}{n_j} > \delta.$$

Boundedness of t_j and $M(\omega', 0) = 0$ also imply that there exists a convergent subsequence $t_{j_k} \to t' > 0$. For any $\epsilon > 0$, we can choose N_ϵ such that for every $k > N_\epsilon$ we have $|\frac{A_i(\omega', n_{j_k} t_{j_k})}{n_{j_k}} - \lambda_i t_{j_k}| < \frac{\epsilon}{2}$ for $i = 1, 2$ and $|t_{j_k} - t'| < \frac{\epsilon}{2(\lambda_1 - \lambda_2)}$, which in turn imply

$$\begin{aligned}
\frac{A_1\left(\omega', n_{j_k} t_{j_k}\right)}{n_{j_k}} - \frac{M(\omega', n_{j_k} t_{j_k})}{n_{j_k}} &= \frac{A_1\left(\omega', n_{j_k} t_{j_k}\right)}{n_{j_k}} - \frac{M\left(\omega', n_{j_k} t_{j_k}\right)}{n_{j_k}} \\
&\quad - (\lambda_1 - \lambda_2)\left(t_{j_k} - t'\right) + (\lambda_1 - \lambda_2)\left(t_{j_k} - t'\right) \\
&> \lambda_2 t_{j_k} - \frac{M\left(\omega', n_{j_k} t_{j_k}\right)}{n_{j_k}} + (\lambda_1 - \lambda_2)t' - \epsilon \\
&> \delta - \epsilon + (\lambda_1 - \lambda_2)t'.
\end{aligned}$$

Similarly, we also get

$$\frac{A_2\left(\omega', n_{j_k} t_{j_k}\right)}{n_{j_k}} - \frac{M\left(\omega', n_{j_k} t_{j_k}\right)}{n_{j_k}} > \delta - \epsilon.$$

Letting $\epsilon \to 0$, we get

$$\min\left\{\frac{X_1\left(\omega', n_{j_k} t_{j_k}\right)}{n_{j_k}}, \frac{X_2\left(\omega', n_{j_k} t_{j_k}\right)}{n_{j_k}}\right\} \geq \delta,$$

which contradicts Lemma 1 and proves (7). This implies

$$\bar{X}_i^n(\omega, t) = \frac{A_i(\omega, nt)}{n} - \frac{M(\omega, nt)}{n} \to \lambda_i t - \min\{\lambda_1, \lambda_2\} t \text{ u.o.c.}$$

for $i = 1, 2$ and every $\omega \in \Omega'$, which concludes the proof. □

The fluid limit (4) does not depend on the matching probability q. This indicates that the users with the lower arrival rate do not accumulate in the system and the system behaves similar to the taxi problem studied in Kashyap [12].

3.2 Diffusion limits

Fluid limits provide useful approximations to determine how queue lengths grow; however, they fail to represent the stochastic fluctuations. To understand the fluctuations of sample paths around the fluid limit, we now focus on diffusion approximations. A direct application of the functional central limit theorem (see, for example, Theorem 5.7 in [6]) on Poisson arrival streams gives

$$\hat{A}_i^n(t) := \frac{A_i(nt) - n\bar{A}_i(t)}{\sqrt{n}} \Rightarrow \hat{A}_i(t), \quad i = 1, 2, \tag{8}$$

where $\hat{A}_i = \sqrt{\lambda_i} B_i$, and $B_i(t), i = 1, 2$, are independent one-dimensional standard Brownian motions. We define the process

$$\hat{X}_i^n(t) := \frac{X_i(nt) - \bar{X}_i(nt)}{\sqrt{n}}, \quad \forall t > 0, \quad n \in \mathbb{N}.$$

Now we are ready to state the diffusion limits for probabilistic matching systems when the matching probability is kept constant.

Theorem 3 *As* $n \to \infty$, $\hat{X}_i^n \Rightarrow \hat{X}_i, i = 1, 2$, *where* $\hat{X}_i$ *is defined as:*

1. *If* $\lambda_1 = \lambda_2$, $\hat{X}_i = \hat{A}_i - \min(\hat{A}_1, \hat{A}_2), i = 1, 2$.
2. *If* $\lambda_1 > \lambda_2$, $\hat{X}_1 = \hat{A}_1 - \hat{A}_2$, $\hat{X}_2 = 0$.

Proof We first consider the case when $\lambda_1 = \lambda_2 = \lambda$. Define $\hat{M}^n(t) := \frac{M(nt)-\lambda nt}{\sqrt{n}}$. Using the Skorohod representation theorem (Theorem 5.1 in [6]), there exist versions of $A_i(t)$, $\hat{A}_i(t)$ and $B_i(t)$, $i = 1, 2$, which we denote $A_i^\circ(t)$, $\hat{A}_i^\circ(t)$ and $B_i^\circ(t)$, $i = 1, 2$, and matching and scaled processes $\hat{M}^{\circ n}(t)$ and $\hat{A}_i^{\circ n}(t)$, $i = 1, 2$ associated with these versions such that $\hat{A}_i^{\circ n}(t) \xrightarrow{\text{a.s.}} \hat{A}_i^\circ(t) = \sqrt{\lambda} B_i^\circ(t)$, $i = 1, 2$. Lemma 1 implies $\min(\hat{A}_1^{\circ n}(t) - \hat{M}^{\circ n}(t),\ \hat{A}_2^{\circ n}(t) - \hat{M}^{\circ n}(t)) \xrightarrow{\text{a.s.}} 0$ u.o.c. Proceeding in the same manner as in the proof of Theorem 2, we get $\hat{M}^{\circ n} \xrightarrow{\text{a.s.}} \min(\hat{A}_1^\circ, \hat{A}_2^\circ)$. Applying the continuous mapping theorem (Theorem 5.2 in [6]), the result follows for $\lambda_1 = \lambda_2 = \lambda$.

When $\lambda_1 > \lambda_2$, let $\tau_n = \inf\{t \geq 0 : A_2(t) \geq n\}$ and define a sequence of random variables $\{\xi_n\}_{n\geq 1}$ such that

$$\xi_n = \begin{cases} 1, & \text{if the } n\text{th arriving user-2 successfully finds a match upon arrival,} \\ 0, & \text{otherwise.} \end{cases}$$

We have $\tau_n \to \infty$ a.s. as $n \to \infty$, and for any $n \geq 1$, $\sum_{n=1}^{A_2(t)} \xi_n \leq M(t)$. Consider a sequence of independent uniform(0,1) random variables $\{U_n\}_{n\geq 1}$. Assuming $0^0 = 1$, we have

$$\begin{aligned}
\mathbb{P}(\xi_n = 0) &= \mathbb{P}\left(U_n < (1-q)^{X_1(\tau_n)}\right) \\
&\leq \mathbb{P}\left(U_n < (1-q)^{A_1(\tau_n) - A_2(\tau_n)}\right) \\
&= \mathbb{P}\left(U_n < (1-q)^{A_1(\tau_n) - n}\right) \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{\{U_n < (1-q)^{A_1(\tau_n)-n}\}} \,\middle|\, A_1(\tau_n)\right]\right] \\
&= \mathbb{E}\left[\left((1-q)^{A_1(\tau_n)-n}\right) \wedge 1\right].
\end{aligned}$$

Next we show that there exist an $N > 0$ and $c > 0$ such that for any $n \geq N$,

$$\mathbb{E}\left[(1-q)^{A_1(\tau_n)-n}\right] < (1-q)^{cn}.$$

For any c_1 such that $1 < c_1 < \frac{\lambda_1}{\lambda_2}$ we have

$$\frac{A_1(t)}{t} - c_1 \frac{A_2(t)}{t} \xrightarrow{\text{a.s.}} \lambda_1 - c_1\lambda_2$$

as $t \to \infty$, i.e., there exists a $T > 0$, such that for $t > T$, $A_1(t) - c_1 A_2(t) > \frac{(\lambda_1 - c_1\lambda_2)t}{2}$ a.s. Since $\tau_n \to \infty$, there exists an $N > 0$ such that for any $n \geq N$, we have $\tau_n > T$ and

$$A_1(\tau_n) - c_1 A_2(\tau_n) = A_1(\tau_n) - c_1 n > \frac{(\lambda_1 - c_1\lambda_2)}{2}\tau_n > 0 \text{ a.s.}$$

Choosing $c = c_1 - 1$ we have $\mathbb{E}[(1-q)^{A_1(\tau_n)-n}] < (1-q)^{cn}$ and

$$\sum_{n=0}^{\infty} \mathbb{P}(\xi_n = 0) = \sum_{n=0}^{\infty} \mathbb{P}\left(U_n < (1-q)^{X_1(T_2(n))}\right) = \sum_{n=0}^{\infty} (1-q)^{cn} < \infty.$$

Using the Borel–Cantelli lemma, $\mathbb{P}(\xi_n = 0 \text{ infinitely often}) = 0$, which in turn implies

$$\hat{X}_2^n(t) = \frac{A_2(nt) - M(nt)}{\sqrt{n}} \xrightarrow{\text{a.s.}} 0.$$

Finally, we have

$$\begin{aligned}\hat{X}_1^n(t) &= \frac{A_1(nt) - M(nt)}{\sqrt{n}} - \frac{(\lambda_1 - \lambda_2)nt}{\sqrt{n}} \\ &= \frac{A_1(nt) - \lambda_1 nt}{\sqrt{n}} - \frac{A_2(nt) - \lambda_2 nt}{\sqrt{n}} - \frac{A_2(nt) - M(nt)}{\sqrt{n}}.\end{aligned}$$

Hence, the result follows from the continuous mapping theorem. □

We conclude that when the matching probability q is kept as a constant in the scaling, it is absent in both the fluid limits and the diffusion limits. Moreover, we can compare our results with those for an $M/M/1$ queue. When the arrival rates in probabilistic matching systems are not equal, the fluid and diffusion limits of the queue length process i behave in accordance with that in an $M/M/1$ queue with arrival rate λ_i and service rate λ_j (see Chen and Yao [6] for more details). When the arrival rates are identical, the diffusion limits are distinct from those of an $M/M/1$ queue, due to the fact that in a probabilistic matching system the next arriving user i is possibly matched immediately upon arrival, which indicates that the accumulation of user j when no user i is present would not be a "waste," unlike the service time generated in an empty $M/M/1$ queue. As a result, rather than having the one-sided regular function of the net-input process, we only have the positive sign of the difference between the arrival processes. We suggest that this diffusion approximation would fit a system which has a relatively high matching probability for each pair of users and thus the probability of an arriving user getting matched increases significantly as the number of users from the other queue grows. However, the underlying assumption above does not hold in those systems which have a very small matching probability for each pair of users, because if q very close to 0, a user is not so likely to find a match upon arrival even when there are many users in the other queue.

4 Fluid and diffusion limits for systems with small matching probability

The matching probability disappears in the fluid and diffusion limits presented in Sect. 3, and this indicates that at most one class of users accumulates in the system and the systems with matching probability $0 < q < 1$ behave very similar to the systems with matching probability 1. However, in many real-world problems the matching

probability q is very small and we need tools that explicitly address the probabilistic nature of the matchings. In this section, we suggest a second type of diffusion approximation which scales q together with the space and time to get a better description of the dynamics of those systems with small matching probabilities.

We often observe that users are impatient and may leave the system without being matched if they cannot match after waiting for some time. We include this factor in the discussion of the queue length process in the new asymptotic regime, adopting a similar approach to that of Ward and Glynn [18], in which the diffusion limit of an $M/M/1+M$ queue with small abandonment rate is provided. We assume that each user has an exponentially distributed abandonment time with rate $\gamma, 0 \leq \gamma < \infty$, independent of others, where γ is of the order of the matching probability and $\gamma \ll \lambda_i, i = 1, 2$. If the abandonment rate γ is significantly greater than the matching probability, as we scale the system the number of matched pairs is negligible compared to the number of users who abandon the system, and the system starts behaving similarly to two independent $M/M/\infty$ queues. Hence, as we scale space, time and the matching probability, we also let the abandonment rate approach zero.

4.1 Fluid limits

Let $X_i^n(t)$ to be the number of class-i users in a probabilistic matching system where class-i users arrive according to a Poisson process with rates λ_i, users abandon the system if they do not match after waiting an exponential time with rate $\gamma^{(n)} = \frac{\gamma}{n}$, $(0 \leq \gamma < \infty)$, and the matching probability is $q^{(n)} = \frac{q}{n}, 0 < q < 1$. Then, we define

$$\bar{X}^{s,n}(t) := \frac{X_i^n(nt)}{n}, \quad \forall t \geq 0,$$

to be the scaled system in this regime. Now our goal is to show that as $n \to \infty$, the scaled system approaches the fluid limit $\bar{X}^s$, which is the unique solution to the following ordinary differential equations (ODEs):

$$\bar{X}_1^s(0) = \bar{X}_2^s(0) = 0, \tag{9}$$

$$\frac{\mathrm{d}\bar{X}_1^s(t)}{\mathrm{d}t} = \lambda_1 \mathrm{e}^{-q\bar{X}_2^s(t)} - \lambda_2 \left(1 - \mathrm{e}^{-q\bar{X}_1^s(t)}\right) - \gamma \bar{X}_1^s(t), \tag{10}$$

$$\frac{\mathrm{d}\bar{X}_2^s(t)}{\mathrm{d}t} = \lambda_2 \mathrm{e}^{-q\bar{X}_1^s(t)} - \lambda_1 \left(1 - \mathrm{e}^{-q\bar{X}_2^s(t)}\right) - \gamma \bar{X}_2^s(t). \tag{11}$$

Define

$$F(x) = \begin{pmatrix} \lambda_1 \mathrm{e}^{-qx_2} - \lambda_2 \left(1 - \mathrm{e}^{-qx_1}\right) - \gamma x_1 \\ \lambda_2 \mathrm{e}^{-qx_1} - \lambda_1 \left(1 - \mathrm{e}^{-qx_2}\right) - \gamma x_1 \end{pmatrix}. \tag{12}$$

Equations (10) and (11) are in the form $\frac{\mathrm{d}x}{\mathrm{d}t} = F(x) = (F_1(x), F_2(x))'$, where $F(\cdot)$ is Lipschitz on the positive quadrant (its component has bounded derivatives), and hence the initial value problem admits a unique solution. We first show that the solution $\bar{X}^s$ is bounded when $\gamma > 0$.

Lemma 4 *Let* $\bar{X}^s = (\bar{X}_1^s, \bar{X}_2^s)$ *be the unique solution to* (9)–(11) *and* $\gamma > 0$*, then*

$$\sup_{0 \le t < \infty} \bar{X}_i^s(t) < \lambda_i/\gamma, \qquad i = 1, 2.$$

Proof For any (x_1, x_2) such that $x_1 \ge \lambda_1/\gamma$, we have

$$F_1(x_1, x_2) = \lambda_1 \mathrm{e}^{-qx_2} - \lambda_2 \left(1 - \mathrm{e}^{-qx_1}\right) - \gamma x_1 < \lambda_1 - \gamma x_1 \le 0.$$

Using (9), this implies that $\bar{X}_1^s(t) \le \lambda_1/\gamma$ for all t. A similar argument also holds for $\bar{X}_2^s(t)$. □

When the matching probability is scaled in such a way that $q^{(n)} \to 0$, the techniques we use to derive fluid and diffusion limits differ from the ones used in Sect. 3. In particular, we appeal to Laplace transform methods where a limiting kernel with the corresponding Laplace transform is identified (see, for example, [8] for a brief review of these methods). For this purpose, we need the Lévy kernel for the Markov process. In this paper, we are dealing with continuous-time pure-jump Markov processes which are time homogeneous. Recall that the Lévy kernel of a pure-jump time-homogeneous Markov process X is defined as

$$\mathbb{P}(X(t + \mathrm{d}t) - X(t) \in [x + \mathrm{d}y] | X(t) = x) = K(x, \mathrm{d}y)\mathrm{d}t.$$

Specifically, the Lévy kernel of $\bar{X}^{s,n}$ is given by

$$\begin{aligned} K^n(x, \mathrm{d}y) := {} & \lambda_1 n \left(1 - \frac{q}{n}\right)^{nx_2} \delta\left((y - x) - \left(\frac{1}{n}, 0\right)\right) \mathrm{d}y \\ & + \lambda_2 n \left(1 - \frac{q}{n}\right)^{nx_1} \delta\left((y - x) - \left(0, \frac{1}{n}\right)\right) \mathrm{d}y \\ & + \left(\lambda_1 n \left(1 - \left(1 - \frac{q}{n}\right)^{nx_2}\right) + \gamma n x_2\right) \delta\left((y - x) + \left(0, \frac{1}{n}\right)\right) \mathrm{d}y \\ & + \left(\lambda_2 \left(1 - \left(1 - \frac{q}{n}\right)^{nx_1}\right) + \gamma n x_1\right) \delta\left((y - x) + \left(\frac{1}{n}, 0\right)\right) \mathrm{d}y, \end{aligned}$$

where $\delta(y)$ is the Dirac delta function. Then, we can define the Laplace transform of the operator $K^n(x, \mathrm{d}y)$ as

$$\begin{aligned} m^n(x, \theta) := {} & \int_{(0,\infty)\times(0,\infty)} \mathrm{e}^{\langle \theta, y \rangle} K^n(x, \mathrm{d}y) \\ = {} & \lambda_1 n \left(1 - \frac{q}{n}\right)^{nx_2} \mathrm{e}^{\frac{\theta_1}{n}} + \lambda_2 n \left(1 - \frac{q}{n}\right)^{nx_1} \mathrm{e}^{\frac{\theta_2}{n}} \\ & + \left(\lambda_1 n \left(1 - \left(1 - \frac{q}{n}\right)^{nx_2}\right) + \gamma n x_2\right) \mathrm{e}^{-\frac{\theta_2}{n}} \\ & + \left(\lambda_2 n \left(1 - \left(1 - \frac{q}{n}\right)^{nx_1}\right) + \gamma n x_1\right) \mathrm{e}^{-\frac{\theta_1}{n}}. \end{aligned} \tag{13}$$

Now, we are ready to state our result for convergence to the fluid limit.

Theorem 5 *For any $\delta > 0$ and $T > 0$,*

$$\limsup_{n\to\infty} n^{-1} \log \mathbb{P}\left(\sup_{0\le t\le T} \left|\bar{X}_i^{s,n}(t) - \bar{X}_i^s(t)\right| > \delta \right) < 0, \tag{14}$$

and as $n \to \infty$,

$$\bar{X}_i^{s,n} \xrightarrow{\text{a.s.}} \bar{X}_i^s \text{ u.o.c.},$$

where $\bar{X}_i^s$, $i = 1, 2$ is the unique solution to the system of ODEs given by (9)–(11).

Proof If $\gamma = 0$, set $\mathbb{S} = \mathbb{R}_{\ge 0} \times \mathbb{R}_{\ge 0}$ and $T^n = T$, otherwise choose $C_i > \lambda_i/\gamma$ for $i = 1, 2$, and set $\mathbb{S} = [0, C_1] \times [0, C_2]$ and $T^n = \inf\{t \ge 0 : \bar{X}^{s,n}(t) \notin \mathbb{S}\} \wedge T$. Then, Proposition 5.1 in [8] implies

$$\limsup_{n\to\infty} n^{-1} \log \mathbb{P}\left(\sup_{0\le t\le T^n} \left|\bar{X}_i^{s,n}(t) - \bar{X}_i^s(t)\right| > \delta \right) < 0 \tag{15}$$

if we can show that the following three conditions hold:

(i) There exists a $\eta_0 > 0$ such that

$$\sup_n \sup_{x\in\mathbb{S}} \sup_{|\theta|\le\eta_0} \frac{m^n(x, n\theta)}{n} < \infty.$$

(ii) $\sup_{x\in\mathbb{S}} |\frac{\partial m^n(x,\theta)}{\partial\theta}|_{\theta=0} - F(x)| \to 0$.
(ii) $\limsup_{n\to\infty} n^{-1} \log \mathbb{P}(|\bar{X}_i^{s,n}(0) - \bar{X}_i^s(0)| > \delta) < 0$.

The third condition is trivially satisfied as we assume that the probabilistic matching system is initially empty and we have $\bar{X}^{s,n}(0) = 0$ for all n. When $\gamma > 0$, the first condition follows as when $x \in \mathbb{S}$ for any $\eta_0 > 0$ and $\theta \le \eta_0$ we have

$$\begin{aligned}
\frac{m^n(x, n\theta)}{n} &= \left(\lambda_1 \left(1 - \left(1 - \frac{q}{n}\right)^{nx_2}\right) + \gamma x_2\right) \mathrm{e}^{-\theta_2} \\
&\quad + \left(\lambda_2 \left(1 - \left(1 - \frac{q}{n}\right)^{nx_1}\right) + \gamma x_1\right) \mathrm{e}^{-\theta_1} \\
&\quad + \lambda_1 \left(1 - \frac{q}{n}\right)^{nx_2} \mathrm{e}^{\theta_1} + \lambda_2 \left(1 - \frac{q}{n}\right)^{nx_1} \mathrm{e}^{\theta_2} \\
&\le (\lambda_1 + \lambda_2)\mathrm{e}^{\eta_0} + \lambda_1 + \lambda_2 + \gamma(C_1 + C_2).
\end{aligned}$$

Similarly, when $\gamma = 0$, the supremum can be bounded by $(\lambda_1 + \lambda_2)\mathrm{e}^{\eta_0} + \lambda_1 + \lambda_2$. To prove the second condition, we write

$$\left.\frac{\partial m^n(x,\theta)}{\partial\theta_1}\right|_{\theta_1=0} = \lambda_1\left(1-\frac{q}{n}\right)^{nx_2} - \left(\lambda_2\left(1-\left(1-\frac{q}{n}\right)^{nx_1}\right)+\gamma x_1\right),$$
$$\left.\frac{\partial m^n(x,\theta)}{\partial\theta_2}\right|_{\theta_2=0} = \lambda_2\left(1-\frac{q}{n}\right)^{nx_1} - \left(\lambda_1\left(1-\left(1-\frac{q}{n}\right)^{nx_1}\right)+\gamma x_2\right).$$

Then it is easy to see pointwise convergence $\frac{\partial m^n(x,\theta)}{\partial\theta}|_{\theta=0} \to F(x)$ and when $\gamma > 0$ the uniform convergence follows from continuity of the functions and the compactness of the underlying set. When $\gamma = 0$, to prove the uniform convergence we directly use the definition. In particular, we need to show that for any $\epsilon > 0$ there exists N such that when $n > N$, we have for any $x \in \mathbb{R}_{\geq 0}$, $|(1-\frac{q}{n})^{nx} - \mathrm{e}^{-qx}| < \epsilon$. First we show that for any $\epsilon > 0$ there exists N_1 and c such that when $n > N_1$ and $x > c$, we have $|(1-\frac{q}{n})^{nx} - \mathrm{e}^{-qx}| < \epsilon$. We know that $\ln(1-\frac{q}{n})^n \to -q$ as $n \to \infty$. For any δ_1 such that $0 < \delta_1 < q$, we can find an N_1 such that for $n > N_1$, we have $x\ln(1-\frac{q}{n})^n < x(-q+\delta_1)$. As a result, letting $c_1 = \frac{\ln\frac{\epsilon}{2}}{-q+\delta_1}$, when $x > c_1$, we have $\ln(1-\frac{q}{n})^{nx} < x(-q+\delta_1) < \ln\frac{\epsilon}{2}$, or equivalently, $(1-\frac{q}{n})^{nx} < \frac{\epsilon}{2}$. Moreover, we know that as $x \to \infty$, $\mathrm{e}^{-qx} \to 0$. We can find a c_2 such that when $x > c_2$, $\mathrm{e}^{-qx} < \frac{\epsilon}{2}$. Letting $c = \max(c_1, c_2)$, the statement follows. Next, due to compactness and pointwise convergence, we know that for $x \in [0, c]$, there exists an N_2 such that for $n > N_2$, we have $|(1-\frac{q}{n})^{nx} - \mathrm{e}^{-qx}| < \epsilon$. Therefore, choosing $N = \max(N_1, N_2)$ we have the uniform convergence for any $x \in \mathbb{R}_{\geq 0}$. As a result, the uniform convergence result for our system when $\gamma = 0$ follows. Therefore, (15) follows from Proposition 5.1 in [8]. When $\gamma = 0$, $T^n = T$ a.s. When $\gamma > 0$, Eq. (15) implies that there is a $\eta > 0$ such that for large enough n

$$\mathbb{P}\left(\sup_{0\leq t\leq T^n}\left|\bar{X}_i^{s,n}(t) - \bar{X}_i^s(t)\right| > \delta\right) \sim \mathrm{e}^{-n},$$

i.e., for any given $\delta > 0$ the probability scales in the order of e^{-n}. Using the Borel–Cantelli lemma, for any $\delta > 0$,

$$\mathbb{P}\left(\sup_{0\leq t\leq T^n}\left|\bar{X}_i^{s,n}(t) - \bar{X}_i^s(t)\right| > \delta \text{ i.o.}\right) = 0,$$

i.e., there exists an Ω' with $\mathbb{P}(\Omega') = 1$ such that for $\omega \in \Omega'$ and $n > N_\delta(\omega)$

$$\sup_{0\leq t\leq T^n(\omega)}\left|\bar{X}_i^{s,n}(\omega,t) - \bar{X}_i^s(\omega,t)\right| \leq \delta. \tag{16}$$

Choose $\delta < (C_i - \lambda/\gamma)/2$ for $\omega \in \Omega'$ and $n > N_\delta(\omega)$, suppose that $T^n(\omega) < T$ and we can reach a contradiction, as $T^n(\omega) < T$ implies that there exists a $t' < T^n$ such that $\bar{X}_i^{s,n}(\omega, t') > C_i - \delta$. This implies that $T^n \overset{\text{a.s.}}{\to} T$. □

When there are abandonments ($\gamma > 0$), the right-hand sides of (10) and (11) involve both e^{-qx} and x terms, which makes it difficult to obtain an analytical solution.

However, when the customers do not abandon the system, the ODEs can be solved analytically. Corollary 6 presents this special case.

Corollary 6 *When $\gamma = 0$, as $n \to \infty$,*

$$\bar{X}_i^{s,n} \xrightarrow{\text{a.s.}} \frac{1}{q} \ln\left(e^{\lambda_1 q t} + e^{\lambda_2 q t} - 1\right) - \mathbb{I}_{\{i=2\}}\lambda_1 t - \mathbb{I}_{\{i=1\}}\lambda_2 t \text{ u.o.c., } i = 1, 2. \quad (17)$$

Proof Setting $\gamma = 0$ and taking the integral of (10) and (11), we see that

$$\bar{X}_1^s(t) + \lambda_2 t = \bar{X}_2^s(t) + \lambda_1 t =: y(t).$$

Then, we have

$$\frac{dy(t)}{dt} = e^{-qy(t)}\left(\lambda_1 e^{\lambda_1 q t} + \lambda_2 e^{\lambda_2 q t}\right)$$

and $y(0) = 0$, which has the unique solution $y(t) = \frac{1}{q}\ln(e^{\lambda_1 q t} + e^{\lambda_2 q t} - 1)$ and the result follows. □

In [3], certain performance measures are proven to be independent of the matching probability q under some control policies. Specifically, [3] considers admission control policies which accept users from class-1 only when $X_1(t) \le X_2(t) + d$ (similarly accept class-2 users only when $X_2(t) \le X_1(t) + d$), where d is a constant, and prove that the difference between long-run average queue lengths of class-1 and class-2 users does not depend on the matching probability q under this admission control policy in Theorem 14. Investigating the proof of this theorem, we see that the global balance equations do not change in a way affecting the result when there are abandonments, and a similar result can be proven. Then letting $d \to \infty$ one might expect the same property for the uncontrolled system. The following corollary confirms this result in the fluid limit and indicates a similar property even under the presence of user abandonments.

Corollary 7 *When $\gamma > 0$, as $n \to \infty$,*

$$\bar{X}_1^{s,n} - \bar{X}_2^{s,n} \xrightarrow{\text{a.s.}} \frac{\lambda_2 - \lambda_1}{\gamma} e^{-\gamma t} + \frac{\lambda_1 - \lambda_2}{\gamma}, \text{ u.o.c.}$$

Proof Applying Theorem 5, $\bar{X}_1^{s,n}(t) - \bar{X}_2^{s,n}(t)$ converges to the unique solution of

$$\frac{dx(t)}{dt} = \lambda_1 - \lambda_2 - \gamma x(t) \quad (18)$$

with initial condition $x(0) = 0$. Using integrating factors, the solution of this first order ODE can be obtained as $\bar{X}_1^s(t) - \bar{X}_2^s(t) = \frac{\lambda_2 - \lambda_1}{\gamma} e^{-\gamma t} + \frac{\lambda_1 - \lambda_2}{\gamma}$. □

Corollary 7 implies that when $\gamma > 0$, the matching probability q does not affect the difference between the numbers of class-1 and class-2 users in the system. As

$t \to \infty$, this difference converges to $\frac{\lambda_1-\lambda_2}{\gamma}$, which coincides with the results of [18] for the $M/M/1+M$ queue with has arrival rate λ_1, service rate λ_2 and abandonment rate $\gamma > 0$.

Next, we analyze the asymptotic behavior of the fluid limits as time goes to infinity. Corollary 6 assumes γ to be 0 and allows us to compare $\bar{X}^s(t)$ with the fluid limits $\bar{X}(t)$, given in Theorem 2. Different from $\bar{X}(t)$, which does not carry any information on the matching probability q, the fluid limits in Corollary 6 depend on q. When t is small, $\bar{X}_i^s(t)$ grows for both $i = 1$ and 2 as q increases. However, as t becomes larger, the influence of the matching probability becomes weaker. Proposition 8 shows that the fluid limits $\bar{X}^s(t)$ converge to $\bar{X}(t)$ as $t \to \infty$.

Proposition 8 *Suppose $\gamma = 0$, then as $t \to \infty$,*

$$\left|\bar{X}_i(t) - \bar{X}_i^s(t)\right| \to 0, \quad i = 1, 2.$$

Proof Without loss of generality, we assume that $\lambda_1 \geq \lambda_2$. Then using Corollary 6 and Theorem 2, we have

$$\begin{aligned}\bar{X}_1^s(t) - \bar{X}_1(t) &= \frac{1}{q}\ln\left(e^{\lambda_1 qt} + e^{\lambda_2 qt} - 1\right) - \lambda_1 t \\ &= \ln\left(e^{\lambda_1 qt} + e^{\lambda_2 qt} - 1\right)^{\frac{1}{q}} - \lambda_1 t \\ &= \ln\frac{\sqrt[q]{e^{\lambda_1 qt} + e^{\lambda_2 qt} - 1}}{\sqrt[q]{e^{\lambda_1 qt}}}.\end{aligned}$$

Since $\lambda_1 > \lambda_2$, we can see that as $t \to \infty$, $|\frac{\sqrt[q]{e^{\lambda_1 qt}+e^{\lambda_2 qt}-1}}{\sqrt[q]{e^{\lambda_1 qt}}}| \to 1$ and this implies that $|\bar{X}_1^s(t) - \bar{X}_1(t)| \to 0$. □

In other words, we can explain the dynamics of a probabilistic matching system in the following way: Without considering the effect of user abandonments, if each pair of users gets harder to match with each other, we observe more users waiting in the system. However, if we run the system long enough, the average numbers of users in the system only depend on the arrival rates. Next we show that for general abandonment rate $\gamma \geq 0$, the fluid limits of the queue length processes converge to a fixed point as $t \to \infty$.

Proposition 9 *If $\gamma > 0$, the fluid limit $\bar{X}_i^s(t) \to x_i^*, i = 1, 2$, as $t \to \infty$, where $x_i^* \in \mathbb{R}$ satisfies the following set of equations:*

$$\lambda_1 e^{-qx_2^*} - \lambda_2\left(1 - e^{-qx_1^*}\right) - \gamma x_1^* = 0, \tag{19}$$

$$\lambda_2 e^{-qx_1^*} - \lambda_1\left(1 - e^{-qx_2^*}\right) - \gamma x_2^* = 0. \tag{20}$$

Proof First, we prove that Eqs. (19) and (20) have a unique solution. Subtracting the second equation from the first one $x_2^* = x_1^* + \frac{\lambda_2-\lambda_1}{\gamma}$, and replacing this into (19), we get

$$\lambda_1 e^{-\frac{q(\lambda_2-\lambda_1)}{\gamma}} e^{-qx_1^*} - \lambda_2 \left(1 - e^{-qx_1^*}\right) - \gamma x_1^* = 0.$$

The left-hand side of the equation is decreasing in x_1^*, equals $\lambda_1 e^{-\frac{q(\lambda_2-\lambda_1)}{\gamma}} > 0$ if $x_1^* = 0$ and goes to $-\infty$ as $x_1^* \to \infty$. Hence, using the intermediate value theorem we conclude that (19) and (20) have a unique solution and $x^* = (x_1^*, x_2^*)$ is the unique fixed point of the system of equations (9)–(11).

When $\lambda_1 \neq \lambda_2$, $\bar{X}^s(t)$ solving (9)–(11) converges to x^* as $t \to \infty$, if we can find a Lyapunov function $V(x)$ with the following properties (see, for example, Strogatz [17]):

1. $V(x) > 0$ for all $x \neq x^*$ and $V(x^*) = 0$.
2. $\frac{dV(\bar{X}^s(t))}{dt} < 0$ for all $x \neq x^*$.

Without loss of generality, we assume that $\lambda_1 > \lambda_2$ and define $V(x) = \lambda_1 - \lambda_2 + \gamma(x_2 - x_1)$. Writing $V(x)$ as $V(x) = \lambda_1 e^{-qx_2} - \lambda_2(1 - e^{-qx_1}) - \gamma x_1 - (\lambda_2 e^{-qx_1} - \lambda_1(1 - e^{-qx_2}) - \gamma x_2)$, we have $V(x^*) = 0$ and $V(x) \neq 0$ for all $x \neq x^*$. Applying Corollary 7 we have $x_1 - x_2 < \frac{\lambda_1-\lambda_2}{\gamma}$ and hence $V(x) > 0$. The second condition follows as

$$\begin{aligned}\frac{dV\left(\bar{X}^s(t)\right)}{dt} &= \gamma\left(\frac{d\bar{X}_2^s(t)}{dt} - \frac{d\bar{X}_2^s(t)}{dt}\right) = \lambda_2 - \lambda_1 + \gamma\left(\bar{X}_1^s(t) - \bar{X}_2^s(t)\right) \\ &= -V\left(\bar{X}^s(t)\right),\end{aligned}$$

which is negative. Therefore, x^* is globally asymptotically stable: For all initial conditions, $\bar{X}^s(t) \to x^*$ as $t \to \infty$.

When $\lambda_1 = \lambda_2 = \lambda$, Corollary 7 implies that $\bar{X}_1^s(t) = \bar{X}_2^s(t)$. Denoting $\tilde{X}(t) = \bar{X}_1^s(t) = \bar{X}_2^s(t)$ and $\tilde{x}^* = x_1^* = x_2^*$ we need to show that $\tilde{X}(t) \to \tilde{x}^*$, $t \to \infty$, where $\tilde{X}(t)$ and $\tilde{x}^*$ satisfy the following equations:

$$\frac{d\tilde{X}(t)}{dt} = 2\lambda e^{-q\tilde{X}(t)} - \lambda - \gamma\tilde{X}(t), \tag{21}$$

$$0 = 2\lambda e^{-q\tilde{x}^*} - \lambda - \gamma\tilde{x}^*. \tag{22}$$

The right-hand side of (22) is a decreasing and (22) is easily seen to have a unique solution. Equation (21) defines a gradient system with potential function $U(x) = \lambda x + \frac{1}{2}\gamma x^2 + \frac{2\lambda}{q}e^{-qx}$, i.e., it can be written as $\frac{d\tilde{X}(t)}{dt} = -\nabla U(\tilde{X}(t))$, where $U(x)$ is a continuously differentiable, single-valued scalar function. Hence, using Theorem 7.2.1 in Strogatz [17] $\tilde{X}(t) \to \tilde{x}^*$, $t \to \infty$. □

The fixed point x^* in Proposition 9 can be thought of as the long-run average of the respective numbers of users of two classes. Hence, the customer abandonment rates in the steady state can be estimated as γx^*. Using the input–output balance, the matching rate, i.e., the average number of users matched in unit time, can also be estimated as $\lambda_i - \gamma x_i^*$ Now, we analyze how x^* behaves for different values of the abandonment rate γ. It is reasonable to expect that x^* should decrease as abandonment rate increases, which

always holds for the user class with the higher arrival rate. However, Proposition 10 shows that for the class with lower arrival rate x^* first increases and then decreases as γ increases.

Proposition 10 *Suppose $\lambda_1 \geq \lambda_2$. Then the long-run average number of class-1 users, x_1^*, decreases as the abandonment rate γ increases, while the long-run average number of class-2 users, x_2^*, increases when*

$$\frac{\lambda_1 - \lambda_2}{\gamma} > \frac{\gamma x_1^*}{q\lambda_1\left(1 - \mathrm{e}^{-qx_2^*}\right) + q\gamma x_2^*}$$

and decreases when the inequality is reversed.

Proof Manipulating Eq. (19) to obtain x_2^*, substituting in Eq. (20) and doing cancellations, we get

$$\ln\left(\lambda_2\left(1 - \mathrm{e}^{-qx_1^*}\right) + \gamma x_1^*\right) = -qx_1^* - \frac{q(\lambda_2 - \lambda_1)}{\gamma} + \ln\lambda_1. \tag{23}$$

Taking the implicit derivative of x_1^* with respect to γ, we obtain

$$x_1^* + \gamma\frac{\mathrm{d}x_1^*}{\mathrm{d}\gamma} + \frac{\gamma}{q}\frac{\mathrm{d}}{\mathrm{d}\gamma}\left[\ln\left(\lambda_2\left(1 - \mathrm{e}^{-qx_1^*}\right) + \gamma x_1^*\right)\right] + \frac{\ln\left(\lambda_2\left(1 - \mathrm{e}^{-qx_1^*}\right) + \gamma x_1^*\right)}{q} - \frac{\ln\lambda_1}{q} = 0. \tag{24}$$

Letting $D_1 = \lambda_2(1 - \mathrm{e}^{-qx_1^*}) + \gamma x_1^*$, $D_2 = \gamma\lambda_2 + \gamma^2 x_1^* + \frac{\gamma^2}{q}$ and substituting Eq. (23) into Eq. (24) to get rid of the logarithm terms, we get

$$\frac{\mathrm{d}x_1^*}{\mathrm{d}\gamma} = \frac{D_1}{D_2}\left(\frac{\lambda_2 - \lambda_1}{\gamma} - \frac{\gamma x_1^*}{qD_1}\right). \tag{25}$$

Since D_1 and D_2 are always positive, when $\lambda_1 \geq \lambda_2$, the right-hand side of Eq. (25) is always negative, and hence as γ increases x_1^* increases. Interchanging x_1^* and λ_1 with x_2^* and λ_2, the right-hand side of Eq. (25) is positive when $\frac{\lambda_1 - \lambda_2}{\gamma} > \frac{\gamma x_2^*}{q\lambda_1(1 - \mathrm{e}^{-qx_2^*}) + q\gamma x_2^*}$ and negative when $\frac{\lambda_1 - \lambda_2}{\gamma} < \frac{\gamma x_2^*}{q\lambda_1(1 - \mathrm{e}^{-qx_2^*}) + q\gamma x_2^*}$. Hence, the conclusion for x_2^* follows. □

Proposition 10 shows that as γ increases, the limiting number of users for the class with lower arrival rate first increases and then decreases, and the limiting number of users for the class with higher arrival rate decreases monotonically, which agrees with the observations in Fig. 3. This behavior can be explained as follows. As the abandonment rate increases, users from both classes tend to abandon the system a lot faster and hence the arriving users from the class with lower arrival rate are less likely

find a match. The decrease in the number of matches is higher than the increase in the abandonments, and as a result, we observe a certain level of accumulation in the limit for users from the class with lower arrival rates.

4.2 Diffusion limits

Now, we move to the discussion of the diffusion limits when the matching probability and the abandonment rate are both scaled to study the fluctuations of the queue lengths around the fluid limit $\bar{X}^s(t)$. We define

$$\hat{X}_i^{s,n}(t) = \frac{X_i^{s,n}(nt) - \bar{X}_i^s(nt)}{\sqrt{n}}, \quad \forall t \geq 0.$$

To prove weak convergence, we again use convergence of generators techniques in [8].

Theorem 11 *Suppose* $\bar{X}^s = (\bar{X}_1^s, \bar{X}_2^s)$ *is the unique solution to the system of ODEs given by* (9)–(11). *Denote*

$$\begin{aligned}
a_1(t) &= q\lambda_2 \mathrm{e}^{-q\bar{X}_1^s(t)},\\
a_2(t) &= q\lambda_1 \mathrm{e}^{-q\bar{X}_2^s(t)},\\
\sigma_1(t) &= \sqrt{\lambda_1 \mathrm{e}^{-q\bar{X}_2^s(t)} + \lambda_2\left(1 - \mathrm{e}^{-q\bar{X}_1^s(t)}\right) + \gamma \bar{X}_1^s(t)},\\
\sigma_2(t) &= \sqrt{\lambda_2 \mathrm{e}^{-q\bar{X}_1^s(t)} + \lambda_1\left(1 - \mathrm{e}^{-q\bar{X}_2^s(t)}\right) + \gamma \bar{X}_2^s(t)},
\end{aligned}$$

and further define

$$\begin{aligned}
z(t) &= \int_0^t \mathrm{e}^{\gamma s}\sigma_2(s)\mathrm{d}B_2(s) - \int \mathrm{e}^{\gamma s}\sigma_1(s)\mathrm{d}B_1(s),\\
z_3(t) &= \mathrm{e}^{\int_0^t a_1(s)+a_2(s)\mathrm{d}s},\\
z_1(t) &= \mathrm{e}^{-\int_0^t a_1(s)+a_2(s)\mathrm{d}s}\left(-\int_0^t z_3(s)a_2(s)z(s)\mathrm{d}s + \int_0^t z_3(s)\mathrm{e}^{\gamma s}\sigma_1(s)\mathrm{d}B_1(s)\right).
\end{aligned}$$

Then we have $\hat{X}^{s,n} \Rightarrow \hat{X}^s$, *where* $\hat{X}^s = (\hat{X}_1^s, \hat{X}_2^s)$,

$$\hat{X}_1^s(t) = \mathrm{e}^{-\gamma t} z_1(t), \tag{26}$$

$$\hat{X}_2^s(t) = \mathrm{e}^{-\gamma t}(z_1(t) + z(t)). \tag{27}$$

Proof Let $\nabla F(x) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} -q\lambda_2 \mathrm{e}^{-qx_1} - \gamma & -q\lambda_1 \mathrm{e}^{-qx_2} \\ -q\lambda_2 \mathrm{e}^{-qx_1} & -q\lambda_1 \mathrm{e}^{-qx_2} - \gamma \end{pmatrix}$,

$$\sigma(x) = \begin{pmatrix} \sqrt{\lambda_1 e^{-qx_2} + \lambda_2\left(1 - e^{-qx_1}\right) + \gamma x_1} & 0 \\ 0 & \sqrt{\lambda_2 e^{-qx_1} + \lambda_1\left(1 - e^{-qx_2}\right) + \gamma x_2} \end{pmatrix},$$

and $\bar{X}^s(t) = (\bar{X}_1^s(t), \bar{X}_2^s(t))'$ be the unique solution to system of ODEs given by (9)–(11). We first show that $\hat{X}^{s,n}(t) \Rightarrow \hat{X}^s(t)$, where $\hat{X}^s(t)$ is treated as a column vector and is the unique solution to the stochastic differential equation

$$d\hat{X}^s(t) = \sigma\left(\bar{X}^s(t)\right) dB_t + \nabla F\left(\bar{X}^s(t)\right) \hat{X}^s(t) dt, \tag{28}$$

starting from $\hat{X}^s(0) = (0, 0)'$, where $B = (B_1, B_2)'$ is a two-dimensional standard Brownian motion. Defining $\mathbb{S}$ as in the proof of Theorem 5 and $F(x)$ as in Eq. (12), the weak convergence follows from Lemma 5.5 in [8], if we can show that the conditions below hold:

(a) $F(x)$ is continuously differentiable on $\mathbb{S}$,
(b) $\sup_{x\in\mathbb{S}} \sqrt{n} |\frac{\partial m^n(x,\theta)}{\partial\theta}|_{\theta=0} - F(x)| \to 0$,
(c) $\frac{\partial^2 m(x,\theta)}{\partial\theta^2}|_{\theta=0}$ is Lipschitz continuous in x on $\mathbb{S}$, where $m(x, \theta)$ is defined by

$$\begin{aligned} m(x,\theta) = &\left(\lambda_1\left(1 - e^{-qx_2}\right) + \gamma x_2\right) e^{-\theta_2} + \left(\lambda_2\left(1 - e^{-qx_1}\right) + \gamma x_1\right) e^{-\theta_1} \\ &+ \lambda_1 e^{-qx_2} e^{\theta_1} + \lambda_2 e^{-qx_1} e^{\theta_2}. \end{aligned}$$

Condition (a) is trivial and condition (b) reduces to showing

$$\sqrt{n}\left(\left(1 - \frac{q}{n}\right)^{nx} - e^{-qx}\right) \to 0,$$

which is elementary calculus, and hence (b) holds as well. Finally

$$\left.\frac{\partial^2 m(x,\theta)}{\partial\theta^2}\right|_{\theta=0} = \begin{pmatrix} \lambda_1 e^{-qx_2} + \lambda_2\left(1 - e^{-qx_1}\right) + \gamma x_1 & 0 \\ 0 & \lambda_2 e^{-qx_1} + \lambda_1\left(1 - e^{-qx_2}\right) + \gamma x_2 \end{pmatrix},$$

which is Lipschitz on $\mathbb{R}^2_{\geq 0}$. Using Lemma 5.5 in [8], $\hat{X}^n \Rightarrow \hat{X}^s$ as $n \to \infty$, where $\hat{X}^s(t)$ is the unique solution to the stochastic differential equation (28). Next we show that (26) and (27) together is the unique solution to (28), which can be expressed as

$$\begin{aligned} d\hat{X}_1^s(t) &= (-a_1(t) - \gamma)\hat{X}_1^s(t)dt - a_2(t)\hat{X}_2^s(t)dt + \sigma_1(t)dB_1(t), \\ d\hat{X}_2^s(t) &= -a_1(t)\hat{X}_1^s(t)dt - (a_2(t) + \gamma)\hat{X}_2^s(t)dt + \sigma_2(t)dB_2(t). \end{aligned}$$

Defining $z_i(t) = e^{\gamma t}\hat{X}_i^s(t)$, $i = 1, 2$, and using integration-by-parts, we obtain

$$\begin{aligned} dz_1(t) &= e^{\gamma t} d\hat{X}_1^s(t) + \gamma e^{\gamma t}\hat{X}_1^s(t)dt \\ &= (-a_1(t) - \gamma)e^{\gamma t}\hat{X}_1^s(t)dt - e^{\gamma t}a_2(t)\hat{X}_2^s(t)dt \\ &\quad + \gamma e^{\gamma t}\hat{X}_1^s(t)dt + e^{\gamma t}\sigma_1(t)dB_1(t) \end{aligned}$$

$$= -a_1(t)z_1(t)\mathrm{d}t - a_2 z_2(t)\mathrm{d}t + \mathrm{e}^{\gamma t}\sigma_1(t)\mathrm{d}B_1(t), \tag{29}$$

and similarly $\mathrm{d}z_2(t) = -a_1(t)z_1\mathrm{d}t - a_2(t)z_2(t)\mathrm{d}t + \mathrm{e}^{\gamma t}\sigma_2(t)\mathrm{d}B_2(t)$. Furthermore, letting $z(t) = z_2(t) - z_1(t)$, we have

$$\mathrm{d}z(t) = \mathrm{e}^{\gamma t}(\sigma_2(t)\mathrm{d}B_2(t) - \sigma_1(t)\mathrm{d}B_1(t)). \tag{30}$$

Solving Eq. (30) directly, we obtain that

$$z(t) = \int_0^t \mathrm{e}^{\gamma s}\sigma_2(s)\mathrm{d}B_2(s) - \int \mathrm{e}^{\gamma s}\sigma_1(s)\mathrm{d}B_1(s).$$

Substituting that $z_2(t) = z(t) + z_1(t)$ into Eq. (29) and moving $z_1(t)$ to the left-hand side, we have

$$\mathrm{d}z_1(t) + (a_1(t) + a_2(t))z_1(t)\mathrm{d}t = -a_2(t)z(t)\mathrm{d}t + \mathrm{e}^{\gamma t}\sigma_1(t)\mathrm{d}B_1(t).$$

Now, multiplying both sides by the integrating factor $z_3(t) = \mathrm{e}^{\int_0^t a_1(s)+a_2(s)ds}$, we get

$$d\left(z_1(t)\mathrm{e}^{\int_0^t a_1(s)+a_2(s)\mathrm{d}s}\right) = z_3(t)\left(-a_2(t)z(t)\mathrm{d}t + \mathrm{e}^{\gamma t}\sigma_1(t)\mathrm{d}B_1(t)\right).$$

As a result,

$$z_1(t) = \mathrm{e}^{-\int_0^t a_1(s)+a_2(s)\mathrm{d}s}\left(-\int_0^t z_3(s)a_2(s)z(s)\mathrm{d}s + \int_0^t z_3(s)\mathrm{e}^{\gamma s}\sigma_1(s)\mathrm{d}B_1(s)\right),$$

$X_1^s(t) = \mathrm{e}^{-\gamma t}z_1(t)$ and $X_2^s(t) = \mathrm{e}^{-\gamma t}(z_1(t) + z(t))$ follow. □

Theorem 11 indicates that if the fluid limit $\bar{X}^s(t)$ is given the diffusion limit can be fully characterized analytically. However, as we have seen in Sect. 4.1, it is not possible to analytically solve the ODEs for the fluid limit when $\gamma > 0$. In the next section, we present numerical experiments to study fluid and diffusion limits presented in this section.

5 Numerical experiments

We now present a numerical analysis of the properties of the fluid and diffusion limits and investigate their performance as approximations to the matching systems. In Sect. 5.1, we use discrete-event simulation to compare the fluid limit presented in Sect. 4.1 with the original matching systems. In Sect. 5.2, we resort to numerical methods for solving ODEs and SDEs to study the fluid and diffusion limits.

5.1 Fluid limits as approximations to the probabilistic matching systems

In Sect. 4.1, we prove that if the matching probability and the abandonment rate goes to zero as we scale time and space, the matching process converges to a fluid limit which can be expressed as the solution of the ordinary differential equations (9)–(11) and these fluid limits converge to the fixed points which solves (19) and (20). This suggests that for "small" matching probabilities and abandonment rates, fluid limits approximate the probabilistic matching systems and the fixed point can be used as an approximation of the average queue lengths of matching queues. Defining L_i to be the long-run average number of class-i users in the system, we now investigate the performance of these approximations and present our results in Tables 1 and 2. The performance measures presented for probabilistic matching systems are obtained using discrete-event simulation.

Table 1 presents a comparison between fluid limits and actual probabilistic matching systems when the arrival rates for both classes are equal. Due to symmetry, the coordinates for the fixed point are equal, i.e., $x_1^* = x_2^* = x^*$, and this value is presented in the third column. In a similar manner, the fourth column presents the average queue length for each class ($L = L_1 = L_2$) as observed in our simulations. The approximation error is given in the fifth column as a proportion of the average queue length. The last two columns correspond to the matching and abandonment rates, respectively. We see that when the matching probability is at most in the order of the abandonment rate normalized by the arrival rate (γ/λ_1), then the fluid approximation provides a good approximation. On the other hand, when the matching probability is significantly greater than the normalized rate, the fluid limit approximation starts to deviate from the real values.

Table 2 presents the performance of fluid limits when $\lambda_2 = 2\lambda_1$. Similar to our results with equal arrival rates, we see that when the abandonment rate normalized by the lower arrival rate (γ/λ_1) is an order of magnitude less than the matching probability, the long-run average number of class-1 users in the system is close to zero and there is significant accumulation of the class-2 users, so that arriving class-1 users find a match upon their arrival. This is an indicator that when the normalized abandonment rate is significantly less than the matching probability, the system behaves similar to the case where the matching probability is equal to one as in the scaling we present in Sect. 3. However, when the matching probability is at most of the order of normalized abandonment rate, then significant numbers of class-1 users wait in the system and the second scaling in Sect. 4 performs well.

5.2 Numerical analysis of properties of fluid and diffusion limits

In Sect. 4, we show that when the matching probability and abandonment rate are scaled to go to zero along with the time and space, the fluid and diffusion limits can be expressed as the unique solutions to some systems of ODEs and SDEs which do not have explicit solutions in general. To gain some insight into the solutions, we study numerical approximations in this section. We use Euler and Euler–Maruyama

Table 1 Comparison of fluid limits and simulation results when $\lambda_1 = 1$ and $\lambda_2 = 1$

γ	q	$x*$	L	$\frac{\lvert x^*-L\rvert}{L}$	M. rate	A. rate
10^{-4}	10^{-4}	3748.2	3743.8	1.19×10^{-3}	125.0	149.796
	10^{-3}	631.88	632.9	1.58×10^{-3}	187.4	25.326
	10^{-2}	68.631	79.6	1.38×10^{-1}	198.4	3.1882
	0.1	6.9245	40.3	8.28×10^{-1}	199.2	1.61286
	0.2	3.464	40.4	9.14×10^{-1}	199.2	1.61724
	0.4	1.7324	40.3	9.57×10^{-1}	199.2	1.61114
10^{-3}	10^{-4}	839.0	838.5	6.62×10^{-4}	32.2	335.5
	10^{-3}	374.8	375.0	3.60×10^{-4}	125.1	149.9
	10^{-2}	63.2	64.2	1.56×10^{-2}	187.2	25.7
	0.1	6.9	14.3	5.19×10^{-1}	197.1	5.7
	0.2	3.4	13.0	7.35×10^{-1}	197.4	5.2
	0.4	1.7	12.8	8.65×10^{-1}	197.4	5.1
10^{-2}	10^{-4}	98.0	98.0	3.62×10^{-4}	3.9	392.2
	10^{-3}	83.9	83.9	3.10×10^{-4}	32.2	335.8
	10^{-2}	37.5	37.5	6.13×10^{-4}	124.9	150.1
	0.1	6.3	7.3	1.29×10^{-1}	185.5	29.0
	0.2	3.3	5.0	3.42×10^{-1}	190.0	20.1
	0.4	1.7	4.2	6.00×10^{-1}	191.5	16.9
10^{-1}	10^{-4}	10.0	10.0	5.01×10^{-4}	0.394	399.1
	10^{-3}	9.8	9.8	2.55×10^{-4}	3.9	392.1
	10^{-2}	8.4	8.4	9.96×10^{-4}	32.2	335.3
	0.1	3.7	3.8	8.78×10^{-3}	124.4	151.3
	0.2	2.4	2.5	4.48×10^{-2}	149.9	100.1
	0.4	1.4	1.7	1.71×10^{-1}	166.1	67.7
2×10^{-1}	10^{-4}	5.0	5.0	3.91×10^{-4}	1.99×10^{-1}	399.5
	10^{-3}	5.0	5.0	6.96×10^{-4}	2.0	396.2
	10^{-2}	4.6	4.6	7.68×10^{-5}	17.8	364.4
	0.1	2.7	2.7	3.16×10^{-3}	93.3	213.6
	0.2	1.9	1.9	1.71×10^{-2}	123.6	152.6
	0.4	1.2	1.3	8.50×10^{-2}	147.7	104.6
2×10^{-1}	10^{-4}	2.5	2.5	2.60×10^{-4}	9.79×10^{-2}	399.7
	10^{-3}	2.5	2.5	4.02×10^{-5}	1.0	398.0
	10^{-2}	2.4	2.4	2.10×10^{-5}	9.4	381.2
	0.1	1.7	1.7	5.83×10^{-4}	62.9	274.2
	0.2	1.3	1.3	5.57×10^{-3}	92.9	214.2
	0.4	9.37×10^{-1}	9.70×10^{-1}	3.43×10^{-2}	122.4	155.3

Table 2 Comparison of fluid limits and simulation results when $\lambda_1 = 1$ and $\lambda_2 = 2$

γ	q	x_1^*	x_2^*	L_1	L_2	M. rate	A. rate
10^{-4}	10^{-4}	1136	11,136	1139.4	11,121	177.1	245.2
	10^{-3}	2.2×10^{-2}	10,000	0.125	9993.6	200.0	200.0
	10^{-2}	1.2×10^{-9}	10,000	1.2×10^{-3}	9987.2	200.0	200.0
	0.1	1.2×10^{-10}	10,000	1.2×10^{-5}	9981.5	200.0	200.0
	0.2	5.9×10^{-11}	10,000	4.0×10^{-6}	9975.1	200.0	200.0
	0.4	2.9×10^{-11}	10,000	2.7×10^{-7}	9992.7	200.0	200.0
10^{-3}	10^{-4}	706.6	1706.7	706.7	1704.9	7.7	584.3
	10^{-3}	113.6	1113.6	113.5	1114.1	58.7	482.3
	10^{-2}	2.16×10^{-3}	1000	3.43×10^{-3}	1000.4	177.3	245.4
	0.1	1.26×10^{-9}	1000	6.72×10^{-6}	1000.1	200 .0	200.0
	0.2	6.28×10^{-10}	1000	5.61×10^{-7}	999.78	200.0	200.0
	0.4	3.14×10^{-10}	1000	1.48×10^{-7}	1000.4	200.0	200.0
10^{-2}	10^{-4}	96.1	196.1	96.1	196.2	7.7	584.6
	10^{-3}	70.7	170.7	70.7	170.6	58.7	482.5
	10^{-2}	11.4	111.4	11.4	111.3	177.1	245.6
	0.1	2.16×10^{-4}	100	4.03×10^{-4}	100.0	200.0	200.0
	0.2	1.15×10^{-8}	100	6.90×10^{-6}	100.0	200.0	200.0
	0.4	3.16×10^{-9}	100	0	100.0	200.0	200.0
10^{-1}	10^{-4}	10.0	20.0	9.9	20.0	7.92×10^{-1}	598.2
	10^{-3}	9.6	19.6	9.6	19.6	7.7	584.6
	10^{-2}	7.1	17.1	7.1	17.1	58.7	482.8
	0.1	1.1	11.136	1.8	11.2	176.7	246.6
	0.2	2.6×10^{-1}	10.3	3.11×10^{-1}	10.3	193.4	212.7
	0.4	2.03×10^{-2}	10.0	4.97×10^{-2}	10.0	198.9	202.1
2×10^{-1}	10^{-4}	5.0	10.0	5.0	10.0	3.97×10^{-1}	599.3
	10^{-3}	4.9	9.9	4.9	9.9	3.9	592.0
	10^{-2}	4.1	9.2	4.2	9.2	33.8	532.3
	0.1	1.4	6.4	1.4	6.4	144.9	310.3
	0.2	5.68×10^{-1}	5.6	6.03×10^{-1}	5.6	176.0	248.3
	0.4	1.31×10^{-1}	5.1	1.86×10^{-1}	5.1	192.5	214.8
4×10^{-1}	10^{-4}	2.5	5.0	2.5	5.0	2.02×10^{-1}	599.7
	10^{-3}	2.5	5.0	2.5	5.0	2.0	596.9
	10^{-2}	2.2	4.8	2.3	4.8	18.3	563.3
	0.1	1.2	3.7	1.8	3.7	105.7	388.6
	0.2	6.84×10^{-1}	3.2	6.96×10^{-1}	3.2	144.2	311.2
	0.4	2.84×10^{-1}	2.8	3.20×10^{-1}	2.8	174.5	251.2

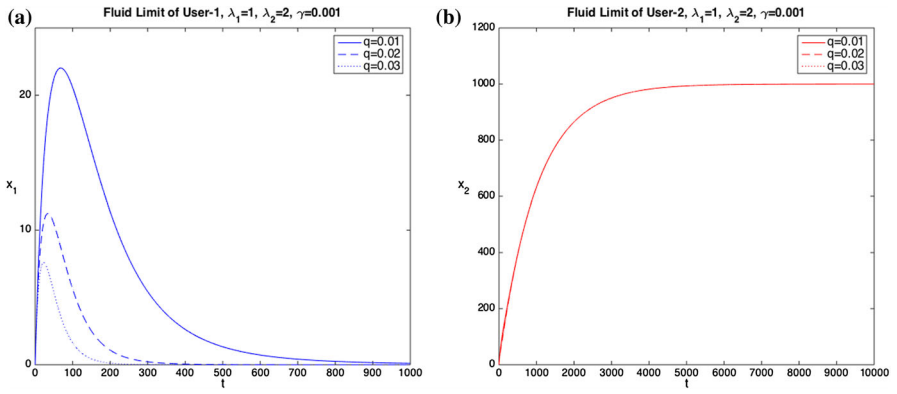

Fig. 1 Fluid limits when $\lambda_1 < \lambda_2$ for various q. **a** Fluid limit of user 1 for different q. **b** Fluid limit of user 2 for different q

methods to obtain numerical solutions of ODEs (9)–(11) and SDEs (28), respectively. (See Kloeden and Platen [13] for more details.)

To study the fluid limit which is the unique solution to the system of ODEs (9)–(11), we apply the Euler method with step size $h = 10^{-6}$. First, we test the effect of the matching probability q on the fluid limits. First we consider the case $\lambda_1 < \lambda_2$ by setting $\lambda_1 = 200$, $\lambda_2 = 400$, $\gamma = 0.5$ and compute the fluid limits for $q = 0.01, 0.02, 0.03$. The results are given in Fig. 1. We observe that for the class with lower arrival rate, the number of users in the system demonstrates a very sharp increase at the beginning and then decreases approaching a limit as t goes to infinity. We see that there is a considerable difference between the number of users corresponding to different matching probabilities for this class. On the other hand, the number of users for the class with higher arrival rate grows monotonically, converging to its supremum as t goes to infinity. Surprisingly, the matching probability does not play a significant role for this class and the fluid limits corresponding to different matching probabilities are very close.

To test the case where $\lambda_1 = \lambda_2$, we performed the same experiment by taking $\lambda_1 = \lambda_2 = 200$. Figure 2a demonstrates that the number of users for both classes increases monotonically as t goes to infinity approaching to the supremum, which is very similar to the behavior of the class with higher arrival rate when the rates are not equal. However, in this case the matching probability has a major effect on the limiting number of users and as q increases the number of users in the system decreases. Also as q gets larger, we see that the number of users increases to its supremum faster and the fluid limit is steeper.

Next we study the effect of the abandonment rate γ on the number of users in the system. In this set of experiments, we set the arrival rates $\lambda_1 = 200$, $\lambda_2 = 400$ and the matching probability $q = 0.01$ and vary the abandonment rate. Figure 3 shows that the shape of fluid limits is not affected by the changes in the abandonment rate, i.e., the number of users for the class with lower arrival rate first increases and then decreases and the number of users for the class with higher arrival rate decreases monotonically. We also see that when there are abandonments the number of users for the class with lower arrival rate does not converge to 0 as t goes to infinity. In

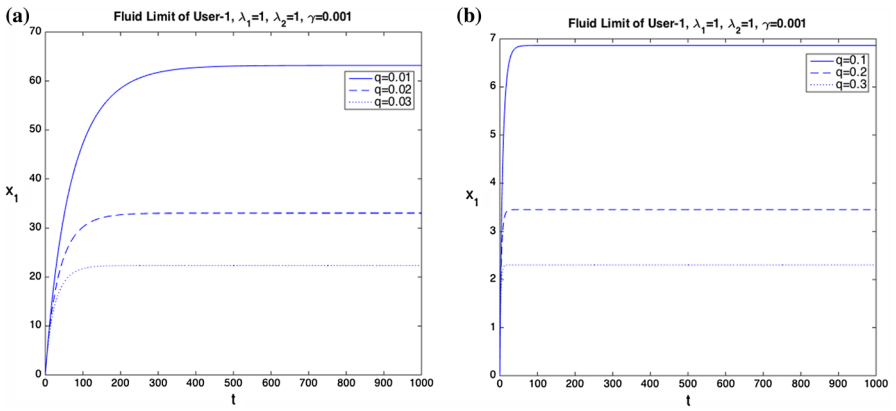

Fig. 2 Fluid limits when $\lambda_1 = \lambda_2$ for various q. **a** Fluid limit of user 1 and 2 for different q. **b** Fluid limit of user 1 and 2 for different q

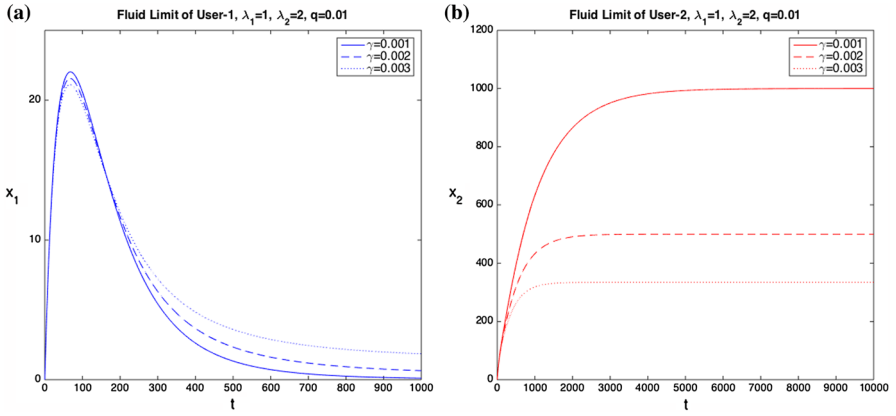

Fig. 3 Fluid limits when $\lambda_1 < \lambda_2$ for various γ. **a** Fluid limit of user 1 for various γ. **b** Fluid limit of user 2 for various γ

agreement with Proposition 10, we see that the limiting number of users for the class with lower arrival rate increases in our experiments as the abandonment rate increases.

Now, we discuss numerical approximation to diffusion limit, which is the unique solution to the system of SDEs (28). In our experiments, we apply the Euler–Maruyama method with the step size $h = 10^{-6}$. We again start with the case when the arrival rates are not equal and set $\lambda_1 = 200$, $\lambda_2 = 400$. Figures 4 and 5 demonstrate some sample paths. We see that there are always significant fluctuations for the class with higher arrival rate. When q is fixed, we see that the changes in γ do not have a major effect on fluctuations. We also see that the fluctuations for the class with lower arrival rate diminish as t increases. However, as q increases, the fluctuations tend to diminish after some time. This is due to the fluid limit approaching zero. Finally, we observe in Fig. 6 that when the arrival rates are equal and set to be $\lambda_1 = \lambda_2 = 200$, both queue length processes keep fluctuating as usual.

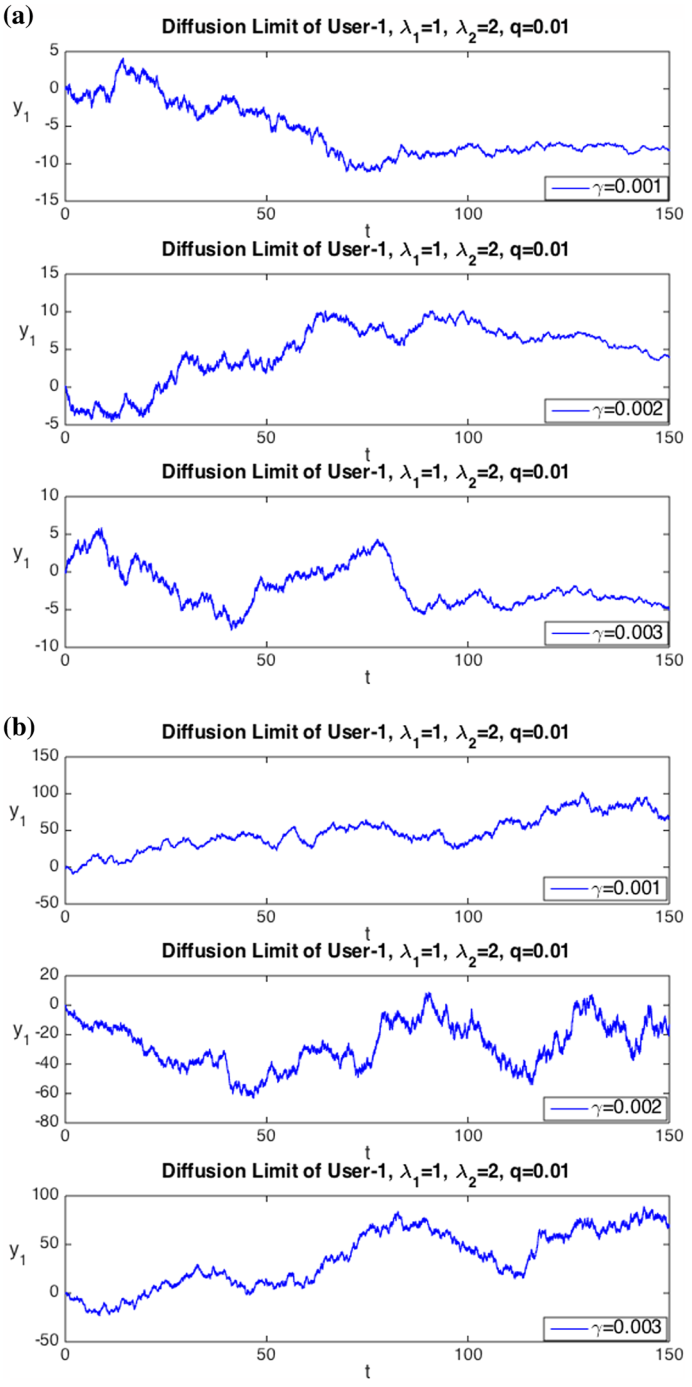

Fig. 4 Diffusion limits when $\lambda_1 < \lambda_2$ for various γ. **a** Diffusion limit of user 1 for various γ. **b** Diffusion limit of user 2 for various γ

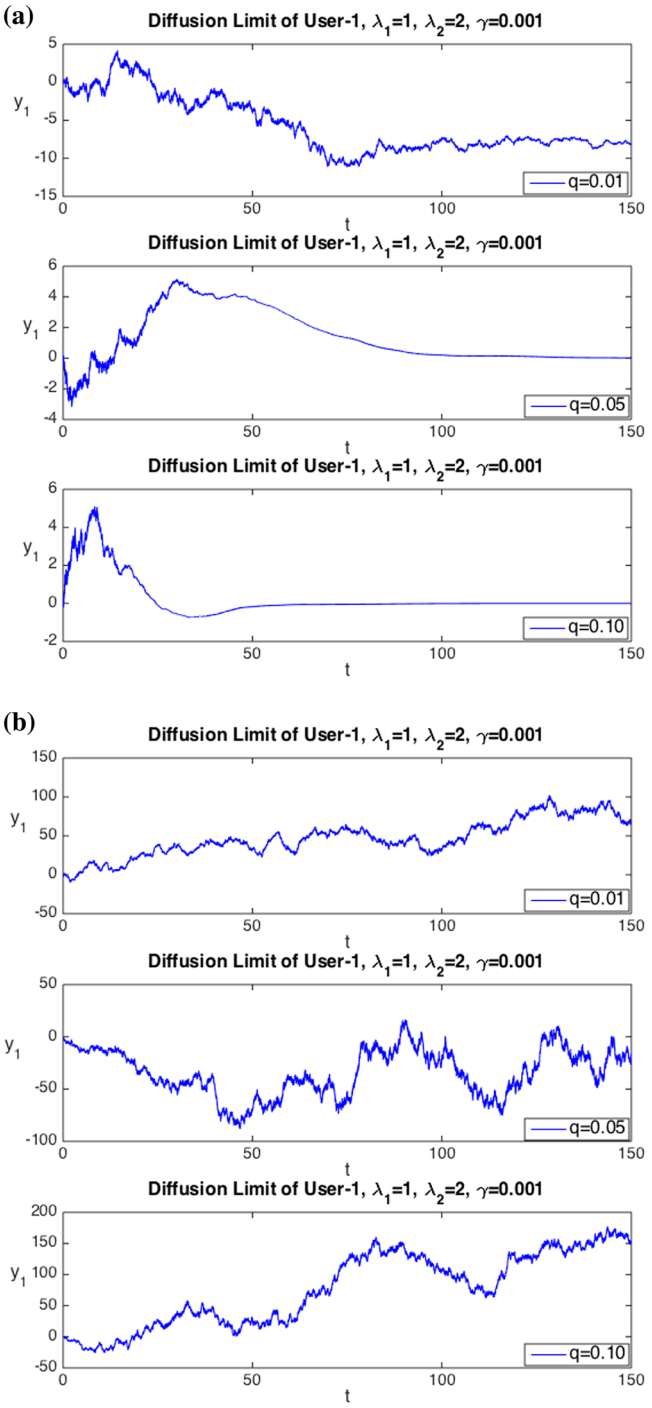

Fig. 5 Diffusion limits when $\lambda_1 < \lambda_2$ for various q. **a** Diffusion limit of user 1 for various q. **b** Diffusion limit of user 2 for various q

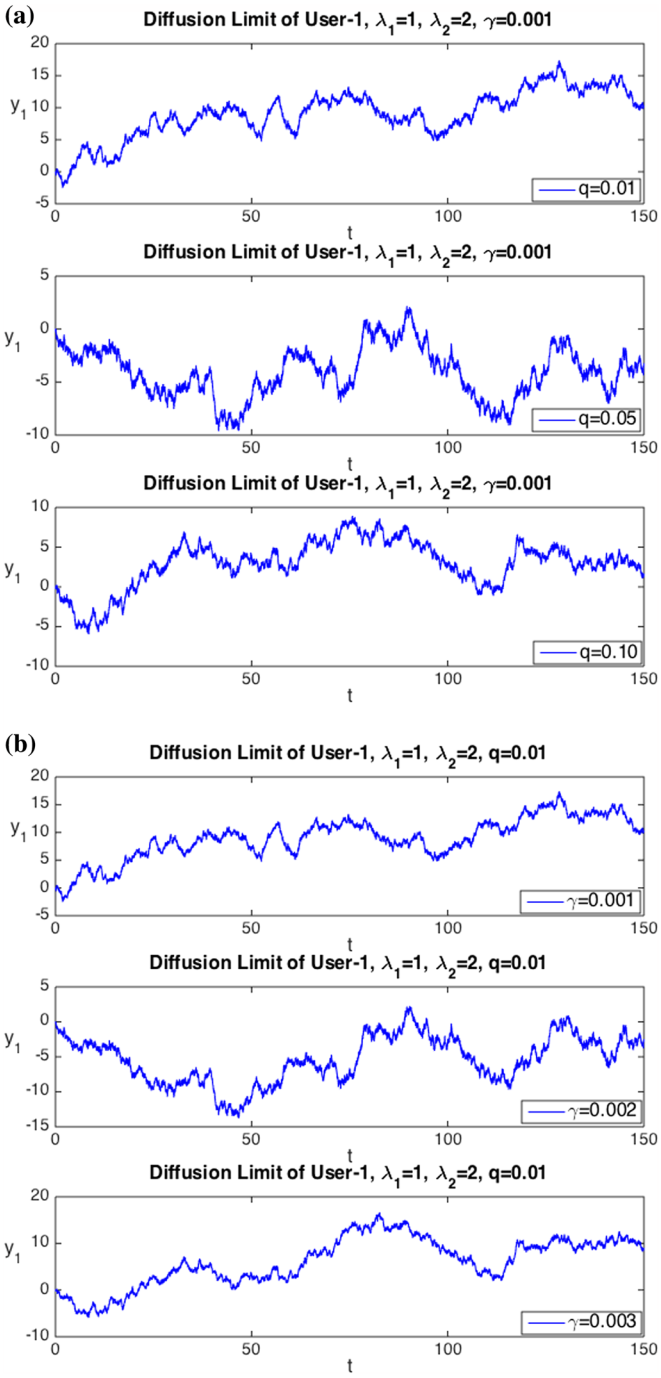

Fig. 6 Diffusion limits when $\lambda_1 = \lambda_2$ for various q and γ. **a** Diffusion limit of user 1 for various q. **b** Diffusion limit of user 1 for various q

6 Conclusion and future work

In this work, we proposed two different scalings to obtain fluid and diffusion approximations to the queue length processes of probabilistic matching systems. For the first approach, the space and time were scaled while the matching probability is kept fixed. Under this scaling, the matching probability q does not play any role in the fluid limit and the minimum of the queue lengths converges to zero. We suggested that this scaling is used when the matching probability is considerably high.

In the second scaling, we addressed the systems in which the probability to match for each pair of users is small. The effect of abandonments was also taken into account, and the matching probability and the departure rate were scaled along with time and space in this regime. The limiting processes enabled us to address the matching probability explicitly. Unfortunately, the resulting system of ODEs cannot be solved analytically in general, although when there are no abandonments it is possible to obtain an analytical solution. In [3], some performance measures were shown to be insensitive to the matching probability under certain admission control policies. Using fluid limits, we showed that the difference between the average queue lengths of different classes of users is also independent of the matching probability. We also analyzed the asymptotic behavior of the fluid limits in this scaling. First, we showed that when the abandonment rate is zero, the two fluid limits, obtained with and without scaling the matching probability, converge to each other with time. We further showed that when there are abandonments, the fluid limits converge to a unique fixed point, which represents the long-run average number of users in the system. Conducting analysis on the fixed point, we revealed that as the abandonment rate increases, the number of users for the class with lower arrival rate first experiences an increase and then decrease, while the number of users for the class with higher arrival rate decreases monotonically.

We also provided extensive numerical results to understand the quality of approximations provided by fluid limits. We saw that if the matching probability is comparable to the normalized abandonment rate or lower our second scaling provides a very good approximation to the real system. However, when the matching probability is significantly higher than the normalized arrival rate, users accumulate in the system and the probability that an arriving lower rate user finds a match approaches one, similarly to what we observe in the first scaling. As analytical expressions are not available for fluid and diffusion limits, we resorted to numerical methods to study the corresponding ODEs and SDEs. We saw that for the class with higher arrival rate, the number of users in the system increases monotonically. On the other hand, the users from the class with lower arrival rate first tend to accumulate in the system and then decrease to a limit as time goes to infinity. This limit is different from zero and increases as the abandonment rate increases, agreeing with our theoretical analysis. This indicates that there are always a significant number of users waiting in the system from both classes.

Probabilistic matching systems exhibit many interesting properties, and we believe the fluid and diffusion limits introduced in this work will be helpful in many directions. First, the approximations introduced here can be used to study the performance of admission control policies which are intractable using exact methods. Another promising research direction is to identify optimal and asymptotically optimal poli-

cies to maximize profit generated by charging users admission fees. The probabilistic matching systems studied in this work can also be extended to include different types of users within each class where each type has a different probability to match with users of other classes. Another possible extension is to consider the situation where each arriving user considers only a subset of users from the other class.

Acknowledgements The work of both authors was supported by the EPSRC grant EP/I017127/1 (Mathematics for Vast Digital Resources).

References

1. Adan, I., Weiss, G.: Exact FCFS matching rates for two infinite multi-type sequences. Oper. Res. **60**(2), 475–489 (2012)
2. Billingsley, P.: Convergence of Probability Measures. Wiley, New York (1999)
3. Büke, B., Chen, H.: Stabilizing policies for probabilistic matching systems. Queue. Syst. **80**(1), 35–69 (2015)
4. Bušić, A., Gupta, V., Mairesse, J.: Stability of the bipartite matching model. Adv. Appl. Probab. **45**(2), 351–378 (2013)
5. Caldentey, R., Kaplan, E., Weiss, G.: FCFS infinite bipartite matching of servers and customers. Adv. Appl. Probab. **41**(3), 695–730 (2009)
6. Chen, H., Yao, D.D.: Fundamentals of Queuing Networks, Performance, Asymptotics, and Optimization. Springer, New York (2001)
7. Dai, J.G., He, S.: Customer abandonment in many-server queues. Math. Oper. Res. **35**(2), 347–362 (2010)
8. Darling, R.W.R., Norris, J.R.: Structure of large random hypergraphs. Ann. Appl. Probab. **15**(1A), 125–152 (2005)
9. Garnett, O., Mandelbaum, A., Reiman, M.: Designing a call center with impatient customers. Manuf. Oper. Res. **4**(3), 208–227 (2002)
10. Gurvich, I., Ward, A.R.: On the dynamic control of matching queues. Stoch. Syst. **4**, 1–45 (2014)
11. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Oper. Res. **29**(3), 567–588 (1981)
12. Kashyap, B.R.K.: The double-ended queue with bulk service and limited waiting space. Oper. Res. **14**(5), 822–834 (1966)
13. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, New York (1999)
14. Liu, X., Gong, Q., Kulkarni, V.G.: Diffusion models for double-ended queues with renewal arrival processes. Stoch. Syst. **5**(1), 1–61 (2015)
15. Mairesse, J., Moyal, P.: Stability of the stochastic matching model. J. Appl. Probab. **53**(4), 1064–1077 (2016)
16. Mandelbaum, A., Momčilović, P.: Queues with many servers and impatient customers. Math. Oper. Res. **37**(1), 41–65 (2012)
17. Strogatz, S.H.: Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering (Studies in Nonlinearity). Westview Press, Boulder (1994)
18. Ward, A.R., Glynn, P.W.: A diffusion approximation for a Markovian queue with reneging. Queue. Syst. **43**(1/2), 103–128 (2003)
19. Ward, A.R., Glynn, P.W.: A diffusion approximation for a $GI/GI/1$ queue with balking or reneging. Queue. Syst. **50**(4), 371–400 (2005)
20. Whitt, W.: Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and their Application to Queues. Springer, Florham Park (2001)